Process requirements for continued scaling of CMOS—the need and prospects for atomic-level manipulation

by P. D. Agnello

Since the advent of the Si-based integrated circuit, ever-increasing function has been available at reduced cost and with reduced consumption of power. This "semiconductor revolution" has been possible because semiconductor devices have the unique feature that as they become smaller they also become faster, consume less power, become cheaper per circuit, and enable more function per unit area of Si. As the basic device approaches atomic dimensions, it is not clear how far scaling can continue, which current processing technologies lack extendibility, and what innovative process technologies will emerge to take their place. Examination of some of the requirements set forth in the *International Technology Roadmap for Semiconductors* (ITRS) [1] will expose some of the process modules that are likely to limit scaling. Future work that might be developed to meet the needs of the CMOS roadmap will then be initiated. Among the processes being developed for future needs are a range of self-limited growth and etching reactions as well as other process steps that take advantage of atomic-level control and manipulation to enable new classes of substrate materials. The requirements that drive such a level of control, as well as the progress and prospects for these new techniques, are discussed in this paper.

Introduction

The objective of this paper is to discuss in some detail the front-end-of-line (FEOL) process steps that are not in common use today but may be required elements for continued progress in the semiconductor industry. The introduction of new, revolutionary process steps such as ion implantation, spacer formation, self-aligned devices, or salicide has enabled progress in the past. New process technologies and materials will emerge to facilitate

©**Copyright** 2002 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/02/$5.00 © 2002 IBM

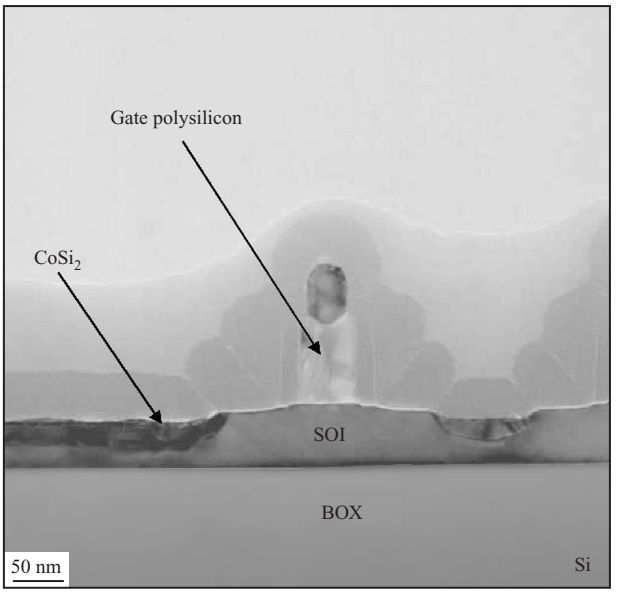


Figure 1

Cross-sectional TEM (XTEM) image of a 90-nm-technology-node device under development, illustrating sub-50-nm physical polysilicon gate.

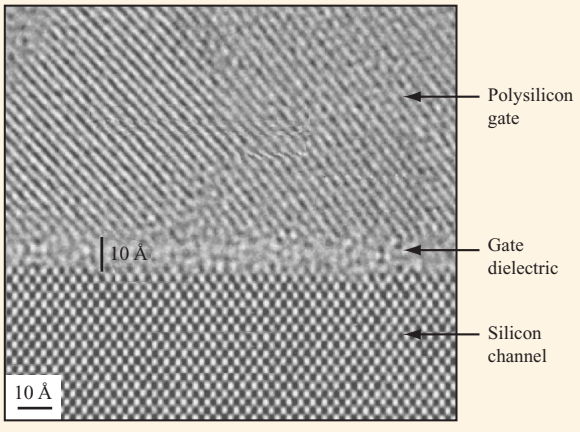


Figure 2

XTEM lattice image of the Si substrate, gate dielectric, and polysilicon for a 90-nm-technology-node device under development.

continued device scaling and performance improvement. Other papers in this special issue focus on device physics as the CMOS device is scaled, and on some novel device structures that may enable continued scaling. Without presuming to know which, if any, novel device structures or materials will take the place of the conventional device structure, this paper examines some of the process issues that affect a selection of different device structures and attempts to determine where new process elements will be indispensable.

The semiconductor revolution began in 1947 with bipolar devices fabricated on slabs of polycrystalline Ge [2]. Single-crystalline materials were later introduced [3], making possible the fabrication of grown junction transistors. Migration to Si-based devices was initially hindered by the stability of the Si/SiO_2 materials system, necessitating a new generation of crystal pullers with improved environmental controls to prevent SiO_2 formation. Later the stability and low interface-state density of the Si/SiO_2 materials system provided passivation of junctions [4] and eventually the migration from bipolar devices to field-effect devices [5] in 1960. By 1968, both complementary metal-oxide–semiconductor devices (CMOS) [6] and polysilicon gate technology that allowed self-alignment of the gate to the source/drain of the device [7] had been developed. These innovations permitted a significant reduction in power dissipation and a reduction of the device overlap capacitance, improving frequency performance and resulting in the essential components of the modern CMOS device. Up until this point, changes involved replacing the whole device structure or a whole element of the device such as the gate electrodes. However, since the advent of the self-aligned, polysilicon-gated transistor, CMOS scaling has been carried out with this fundamental device at the core. Subsequent changes have taken place at the periphery of the device. **Figure 1** is a high-resolution transmission electron microscopy (TEM) image of a device under development for the 90-nm-technology node. The device is built on thin silicon-on-insulator (SOI) substrates and has self-aligned silicide and dual spacers; however, except for these additions, the details of the junction fabrication, and the absolute dimensions of the device, it is remarkably similar to devices fabricated more than 30 years ago.

The first process module that will reach the atomic limit is the gate dielectric module. **Figure 2** is a high-resolution TEM lattice image of a silicon oxynitride gate dielectric film under development for the 90-nm-technology node. The dielectric is only a few monolayers in thickness. Continued scaling will require a reduction in this thickness by 70% every subsequent generation in order to maintain short-channel control as the physical gate length is scaled. Will a move to a different device structure, still employing the Si/SiO_2 interface, be more likely than a change to high-k dielectrics in the materials set? Both types of changes (new device structure and new materials set) have occurred in the past, but the change in materials set for a fixed device structure (Ge bipolars migrating to Si bipolars) occurred only before the Si/SiO_2 system was well understood. Subsequent changes to that device structure

have occurred, but the Si/SiO_2 interface has played a pivotal role in all of them. For example, planar SiO_2-passivated bipolars were supplanted by Al-gated SiO_2 FETs, and then by dual-workfunction polysilicon-gated FETs. More recently, Si CMOS devices have been improved by fabrication on SOI substrates. However, all of those devices revolved around the Si/SiO_2 interface, which suggests that migration to high-*k* gate dielectrics or other materials for the channel of the FET will be very difficult. For this reason some workers [8] have suggested that radical new devices based on quantum effects and interactions may be required, because such devices could take advantage of the Si technology infrastructure and the well-understood SiO_2/Si interface [9]. Whether the industry pursues continued CMOS scaling with different materials or a new class of devices utilizing the Si/SiO_2 materials system, there will be a need to control processes to atomic levels, which is the subject of discussion in this paper.

Requirements for thin-film deposition

As devices are scaled, there will be a need to control the thickness of thin-film depositions to the atomic-layer scale. **Figures 3** and **4**, respectively, are plots of gate dielectric and sidewall spacer requirements as defined by the ITRS roadmap [1]. The requirements for gate dielectric equivalent-oxide thickness (EOT) tolerance approach 0.02 nm by the year 2014. Since scaling of the conventional device will necessitate that the gate dielectric move away from SiO_2-based films by approximately 2005, the physical tolerance requirements will be relaxed by the ratio of the dielectric constant of the new material to that of SiO_2. In the case of the sidewall spacer, thickness-control requirements will approach the 2-nm level by 2014. Given that today's gate dielectric tools can achieve tolerances of less than 0.01 nm for 1.5-nm films, these requirements may not appear to be too stringent on the basis of current capability. However, that tolerance is currently achieved by thermally grown films whose depositions are highly sublinear with growth time and are thus much more reproducible than chemical-vapor-deposited (CVD) films. Many of the films in FEOL processing are deposited by CVD or plasma-enhanced CVD (PECVD), and achieving the required film thickness control will be a challenge for those techniques in the future.

A simplified model shows that the CVD processes are governed by one of two mechanisms: kinetically controlled or mass-transport-limited. At low growth temperatures, the rate of growth is low and there is an overabundance of reactant species. The growth rate is determined by the rate of thermally activated surface reactions with activation energies ranging from 1 to 4 eV per molecule; it can be dependent on the growth surface and the crystal orientation, in the case of epitaxy. The high activation

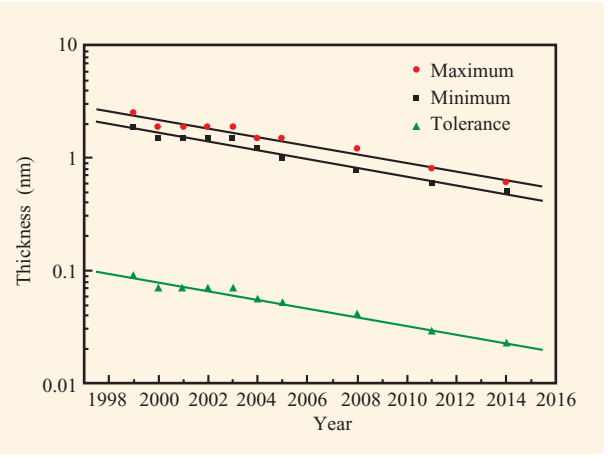


Figure 3

Minimum and maximum projected requirements for the gate dielectric equivalent-oxide thickness and the projected tolerance as defined by the ITRS [1].

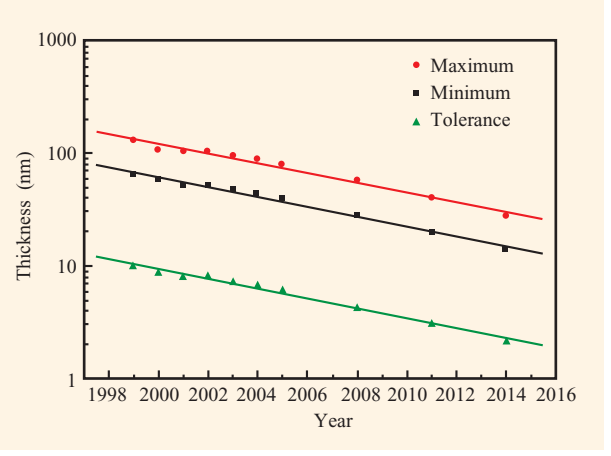


Figure 4

Minimum and maximum projected requirements for the sidewall spacer thickness and the projected tolerance as defined by the ITRS [1].

energies in this growth regime make temperature control extremely critical to growth uniformity. As the temperature is increased, the rate of surface reactions increases to the point at which the supply of reactants or the transport of reaction byproducts away from the surface limits the growth rate.

The mass-transport-limited regime is characterized by a lower activation energy, typically of a few tenths of an eV per molecule, due to the temperature dependence of the

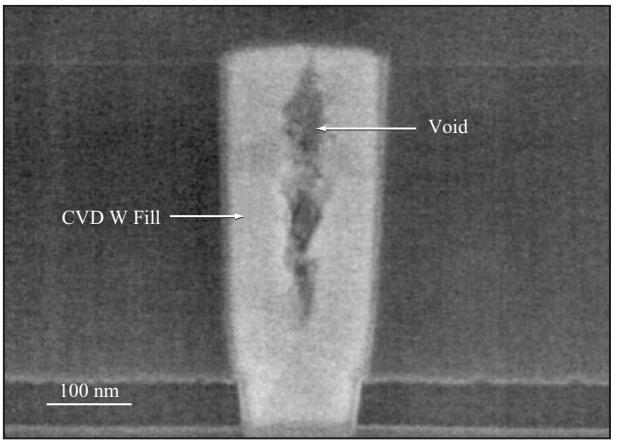


Figure 5

SEM image of a CVD W-filled contact hole for 130-nm-technology-node CMOS. A CVD TiN liner was deposited on a thin i-PVD seed layer preceding the W CVD fill. A pronounced void in the fill is observed.

gas-phase diffusion constant. Since the supply of reactants at the wafer surface is limited in this case, the local growth rates across a wafer can vary considerably because of macroscopic depletion of reactant in the reaction chamber or along the wafer, as well as microscopic depletion due to the local density of patterns on the wafer. Models have been developed that take into account microloading over three different length scales: wafer level, device level, and an intermediate scale, in order to improve predictive capability and design better reactors [10]. Even so, it is not likely that current CVD processes can be controlled to the desired precision, over all structures, across ever-larger wafers.

Both spacer sidewall film depositions and selective raised source/drain contacts are good examples of structures in which CVD technology has reached or will approach the limits of conventional techniques. Sidewall spacer deposition thickness is currently a function of the gate pattern density because the growth rate is limited by the reactant diffusion rate. This results in a slower deposition rate in the local vicinity of high-pattern-density gate features due to the increased surface area for deposition with a fixed supply of reactant. Selective epitaxial growth processes such as those proposed to achieve raised source/drain contacts are another process expected to be difficult to control adequately. At higher growth temperatures, there are significant problems with the control of local growth rate because reactants adsorb on masked regions in a selective process and migrate over the surface, resulting in locally enhanced growth rates. At lower growth temperatures, growth on n-type and p-type regions varies significantly because of the kinetically controlled growth process.

Contact liners are another example in which the limits of conventional physical vapor deposition (PVD) have already passed, and those of CVD are rapidly approaching. Contact liners have progressed from PVD to collimated PVD to ionized PVD (i-PVD) and finally to CVD in order to address the concern of liner pinch-off of contact features. Coverage of the sidewalls at the bottom of the hole can be limited to a few percent of that deposited on the upper rim of the contact hole because of the angular distribution of sputtered species from the target.

Collimated and ionized PVD techniques provide a more directional flux at the substrate surface in order to provide more material to the bottom of high-aspect-ratio contact holes. In spite of these advances in PVD technology, some applications have moved to CVD liner processes. Even conventional CVD techniques will encounter limitations in small contact holes as the flux of reactants and byproducts into and out of small, high-aspect-ratio contact holes becomes restricted. **Figure 5** is a cross-sectional SEM of a filled contact hole for 130-nm-technology-node CMOS, utilizing a very thin i-PVD seed layer followed by a CVD TiN liner and CVD W fill processes. A seam in the fill is observed. Other examples of process steps that are subject to keyholes and seams are shallow-trench isolation (STI) fill and interlayer dielectric (ILD) dielectric deposition that precedes the first contact level. Current approaches to these problems have been to reduce the pressure in the reactor in order to increase the diffusivity of molecules. The pressure can only be reduced by a certain amount before the flux of reactant species becomes too small to be practical, and in PECVD reactions, charging problems can result from plasma damage at low pressures. Sequences of deposition and etching (dep-etch processes) provide some improvement in filling high-aspect-ratio features, but uniformity over all pattern factors is not easily achieved in a process that involves balancing competing reactions.

Extension junction and contact resistance

Another concern for continued scaling of CMOS devices is external resistance (R_{ext}). As devices are scaled, the on resistance, R_{on}, of the intrinsic device is reduced, and it is important to keep the parasitic source/drain resistance a small fraction of R_{on} in order to maintain good transconductance and overall performance at the device terminals. **Figure 6** is a plot of R_{on} of high-performance n-FET devices by year of volume manufacturing buildup for high-performance logic technologies. A similar plot can be made for p-FETs. The historical trend has been to keep the total external resistance, R_{ext}, to about 10% of R_{on} in order to maintain good performance. With the approach of the 90-nm-technology node, n-FET R_{ext} less

than 100 Ω-μm will be required. Further scaling for the 60-nm node and beyond will be a considerable challenge, since it requires R_{ext} significantly less than 100 Ω-μm.

The total external resistance equals

$$R_{ext} = 2(R_{ac} + R_{sp} + R_{sh} + R_{co}), \quad (1)$$

where R_{ac} is the accumulation layer resistance, R_{sp} is the spreading resistance of the junction, R_{sh} is the sheet resistance of the junction, and R_{co} is the contact resistance between the contact metal and the silicon [11]. The contact resistance R_{co} is a function of the specific contact resistivity of the silicide/silicon interface ρ_c, as well as the sheet resistance $\rho_{\square}$ of the doped silicon, which determines the geometric flow of current. In the case where $\rho_{\square}$ is low with respect to ρ_c, the current can flow over the entire silicide length L_s into the silicon. In this case R_{co} is simply

$$R_{co} = \rho_c / L_s W \quad (2)$$

for a device of width W. However, for the case in which silicide consumes much of the deep junction or for SOI on thin Si, such that the $\rho_{\square}$ under the silicide is high, the effective length for current flow is reduced, and the contact resistance saturates at

$$R_{co} \approx \sqrt{(\rho_c \rho_{\square})}/W. \quad (3)$$

The length L_s essentially drops out of this expression, and current flow can be thought of as flowing out of an effective length ℓ' [11]. The specific contact resistivity has approximately the dependence

$$\rho_c \approx \exp\left[(4\pi\sqrt{(\varepsilon_s m^*)}/h)(\phi_b/\sqrt{N_d})\right], \quad (4)$$

where the two important parameters which can be controlled are the substrate doping concentration N_d and the barrier height at the silicide/silicon interface ϕ_b.

The sheet-resistance term of the external resistance is dominated by the extension sheet-resistance component in present devices. Approximately,

$$R_{sh} = \rho_e L_e/(X_j W), \quad (5)$$

where ρ_e is the specific resistivity of the extension junction, L_e is the length of the extension junction, and X_j is its depth. L_e is typically less than the spacer width L_{sp} because the thermal cycle and lateral straggle of the extension junction are less than those of the deep junction. If the values of ρ_e, X_j, and L_{sp} are taken from the ITRS roadmap and it is assumed that $L_e \sim 1/2\ L_{sp}$, the value of R_{sh} is currently from 40 to 10 Ω-μm and scales lower with continued scaling. However, in order to keep R_{sh} a small fraction of R_{ext}, continued scaling of ρ_e will be required even as the junction depth is reduced.

Figures 7(a) and **7(b)**, respectively, are plots of the junction sheet resistance vs. junction depth from the requirements in the ITRS roadmap for p-FETs and n-FETs. The sheet resistance vs. junction depth for ideal boxlike profiles at constant dopant concentration is plotted for reference. It is clear that active dopant levels exceeding 1×10^{21} cm^{-3} are required. The performance of current ultralow implant energy plus optimized RTA-based junctions is also shown in the figures [12, 13], indicating that the p-FET targets will be difficult to achieve after the 130-nm node, while those for the n-FET will be difficult after the 60-nm node.

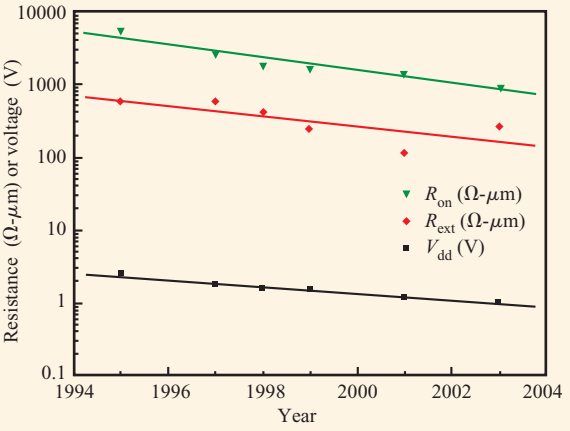


Figure 6

R_{on}, R_{ext}, and supply voltage (V_{dd}) trends by year of technologies in or nearing production.

Two other important components of the external resistance are the spreading and accumulation resistance of the contact from the extension into the inversion layer. The ITRS roadmap assumes a junction abruptness, X_j, as shown in **Figure 8**. The improvement in extension junction abruptness is needed for three reasons: improved short-channel effect, reduced spreading resistance, and reduced accumulation resistance. With present implanted plus RTA junction technology, it is possible to achieve slopes better than 5 nm/decade for 1-keV As implants [13], but scaling of X_j is required in the ITRS because by current techniques the lateral and vertical junction depths are coupled, making it impossible to meet the X_j and $\rho_{\square}$ requirements with the same process.

To summarize the requirements placed on extension junction design by reducing R_{ext}, reduced sheet resistance, R_{sh}, requires heavy doping, but shallow junction depth degrades the sheet resistance [see Equation (5)]. Reduced contact resistance, R_{co}, also requires heavy doping, but reduced junction depth compromises the $\rho_{\square}$ term in Equation (3) while reducing the amount of Si that can be consumed for silicide formation without having degraded contact resistance due to the silicide intersecting

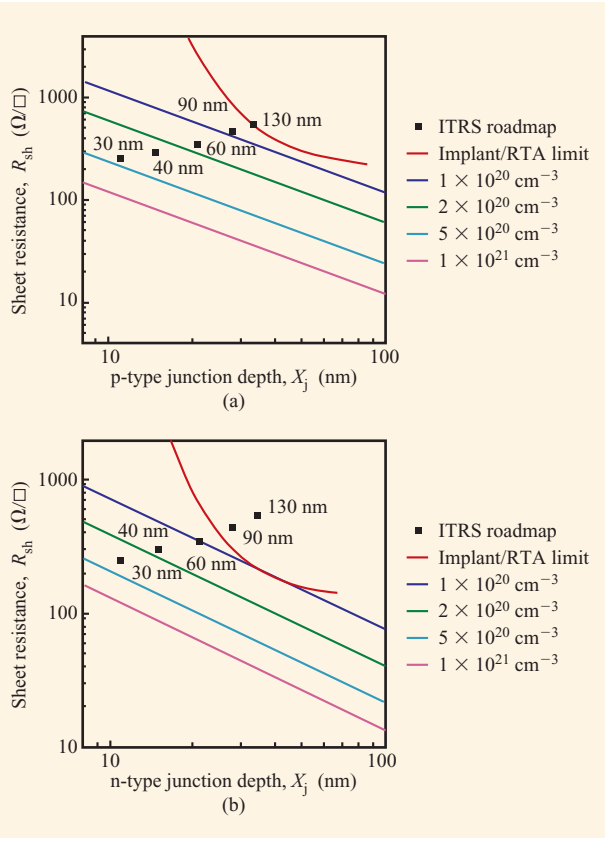


Figure 7

Sheet resistance vs. extension junction depth for (a) p-FET and (b) n-FET. The squares represent the ITRS requirements for the 130-nm through 30-nm nodes. Since the specified requirements vary by about a factor of 2 in sheet resistance and by a factor of 3 in X_j, the mean values were plotted. The curves represent the predicted values for boxlike profiles of dopant concentrations 1×10^{20}, 2×10^{20}, 5×10^{20}, and 1×10^{21} cm^{-3} and are plotted for reference.

the junction at low carrier concentration [Equation (4)]. Reduced spreading and accumulation resistance, R_{sp} and R_{ac}, require heavy doping with very abrupt profiles. For junctions formed by implantation and RTA, profile abruptness can be degraded with deeper junctions, thus putting minimization of R_{ac} and R_{sp} at odds with minimizing R_{sh} and R_{co}. Laser annealing [14] has been utilized to achieve ultrashallow junctions with high activation due to the high solubility of dopants at or near the melting point of Si. However, it is not clear whether this technique can be integrated into the conventional process flow, because the laser pulse is absorbed in the gate conductor and results in damage. Utilizing laser annealing with a replacement gate flow may circumvent this problem, but care will be required to prevent deactivation of the dopants with subsequent processing.

The raised source/drain technique, in which epitaxial Si is deposited selectively after the extension junctions are in place, can help to decouple the X_j term from the ρ_e and $\rho_{\square}$ terms and can also allow intersection of the silicide with the doped Si interface at a higher value of N_d, thereby reducing ϕ_b as well. There are several variants to the sequence of extension junction formation, deep junction formation, and selective epitaxy that can be utilized [15], but most require that a shallow, abrupt, and highly doped junction first be formed by implantation and annealing, and care must be taken to ensure that the epitaxial growth does not deactivate the dopants or lead to transient-enhanced diffusion. Therefore, growth at low temperatures is desired, but growth rates on n-FET and p-FET regions of the CMOS circuit can be quite different because the growth rate is kinetically controlled by surface reactions.

Atomic layer deposition

As a result of the limitations of conventional PVD, CVD, and RIE techniques, processes with atomic-layer control are required for continued technology scaling. By using a sequence of self-limiting surface reactions, atomic-layer control of thin-film depositions was disclosed by Suntola and Antson in their 1977 patent [16]. The mechanism originally proposed was for the deposition of compounds, but it has since been extended to elemental films. In the case in which the film is grown epitaxially on a crystalline substrate, the technique is known as atomic layer epitaxy (ALE); otherwise it is known as atomic layer deposition (ALD). ALD and ALE can now be used to deposit a wide range of materials. Commercial tools are currently available for the ALD of high-k gate dielectrics and W liner materials, but ALD may be useful for many other process steps. It could enable controlled fabrication of many other parts of the device to tight tolerances if tooling and processes for the ALD of other materials, such as Si for raised source/drain contacts and SiO_2 and Si_3N_4 for spacer materials, could be brought to a production-worthy level. Examining the status of ALD and ALE deposition of a range of materials will give a sense of the probability of extending ALD beyond W and high-k gate dielectrics. Further, the prospects for using self-limiting surface reactions for material removal as well as deposition will be explored.

The first demonstration of this technique by Ahonen, Pessa, and Suntola in 1980 [17] was for the deposition of ZnTe films, where molecular beams of the pure elements were used. Specifically, it was demonstrated that layer-by-layer growth of II–VI compounds can be achieved by taking advantage of the fact that the equilibrium vapor pressures of the II–VI compounds and their pure constituents are very different, and that the adsorption coefficients of the impinging atoms depend strongly on

the atomic species of the uppermost layer of the film. If, in contrast to conventional MBE, in which fluxes of multiple species are provided simultaneously, only one elemental flux is provided, one monolayer will be adsorbed but additional species will not be strongly adsorbed because of the lack of formation of a II–VI chemical bond. By making use of the fact that at temperatures below that for congruent evaporation, the vapor pressures of the elements exceed that of the compound, excess reactant atoms can be made to re-evaporate before a beam of the next species is allowed to react with the surface.

As a result, ALE provides a process in which films are largely insensitive to temperature and the flux of reactant species. This is in contrast to conventional molecular beam epitaxy of compound semiconductors, in which the growth rate and the film quality are strong functions of the temperature and fluxes of the reactant species. In ALE reactants are provided in sequential pulses: -A-B-A-B- etc. As long as at least one monolayer of reactant can adsorb in each pulse, and there is sufficient time between pulses to desorb excess reactant, layer-by-layer growth is achieved. The dependence of growth rate on temperature is also reduced. For temperatures below the temperature for congruent evaporation, but still high enough for sufficient surface mobility of adsorbed species and desorption of excess species, each A-B sequence provides one monolayer of compound growth.

Temperature-insensitive ALE growth of ZnTe is obtained between 593 and 673 K, with a growth rate of approximately one ZnTe monolayer per -A-B- cycle [17]. Above 673 K, the growth rate drops to near zero as evaporation of the compound occurs. ALE for the case in which elemental molecular beams are not used was demonstrated in 1981 [18]. Sequential exchange reactions were used to obtain layer-by-layer growth of Ta_2O_5 and ZnS films from a combination of pulsed molecular beams of the metal chloride from effusion ovens interleaved with gas pulses of either H_2O or H_2S to complete the reaction cycle. To study the exchange reactions taking place, Auger spectroscopy of the growth surfaces was performed after each pulse. In both cases, the weakly chemisorbed Cl present on the surface after the metal chloride adsorption is completely removed by the adsorption of either H_2O or H_2S. This reaction produces HCl vapor and either Ta_2O_5 or ZnS films, respectively, with no traces of Cl present. After multiple pulses, films >300 nm thick can be produced that are stoichiometric and, in the case of ZnS, crystalline and highly oriented along the growth axis.

This technique was demonstrated in the III–V materials system for the growth of GaAs in what might be best described as a gas-source MBE apparatus [19]. Arsine and trimethylgallium were used as the reactants, introduced by sequential pulsing, with the chamber evacuated between pulses. While self-limited monolayer-by-monolayer growth

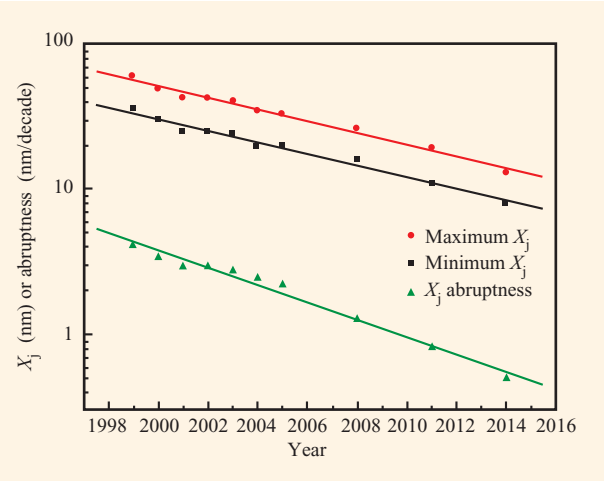


Figure 8

Minimum and maximum projected requirements for junction depth and the projected value for junction abruptness from the ITRS [1].

was achieved under these conditions, poor-quality material was produced, with mobilities of ~100 cm^2/V-s and carrier densities greater than $1 \times 10^{18}/cm^3$. Later that same year the technique was extended to CVD growth at more conventional operating pressures for both GaAs and AlAs [20]. The same reactants were used as mentioned above [19], but in an atmosphere of hydrogen gas, as for growth by conventional organometallic CVD. Layer-by-layer growth is achieved by placing the substrate on a rotating susceptor and passing it under different reactant streams separated by gas curtains. The material quality obtained by this method is much better, with 77 K photoluminescence peak widths of 11 meV obtained. Later, GaAs was grown by similar means by other workers, and room-temperature electron mobilities exceeding 5500 cm^2/V-s were obtained, with residual carrier densities less than $1 \times 10^{15}/cm^3$ [21]. Considering these separate investigations reveals an important point; in order to ensure good materials quality by ALE, it is important to maintain a passivated surface during growth. While high-quality films are obtained by conventional MBE, it is more difficult to obtain them by MBE in the ALE mode, where the growth has to be interrupted to allow desorption of reaction byproducts or switching of species. At this step, significant incorporation of impurities can occur. In contrast, ALE by CVD means is potentially more manufacturable, at least in the case in which a suitable -A-B- reaction sequence can be found, as in the growth of GaAs and AlAs. CVD-based ALE has not been as straightforward for the Si system.

Extension of the ALE growth technique to an elemental semiconductor such as Si presents additional challenges.

Early work utilized adsorption of Si_2H_6 at cryogenic substrate temperatures followed by laser irradiation by an ArF excimer source at 193 nm and evacuation to desorb reaction byproducts [22, 23]. Deposition occurs only on the Si surface and not on SiO_2, unless the adsorption time is substantially longer than that used for deposition on Si. SiH_2Cl_2 and H_2 have also been used as precursors [24]. At intermediate temperatures in the vicinity of 850°C, SiH_2Cl_2 decomposes to $SiCl_2$ on the surface with the desorption of H_2. Subsequent evacuation of the reactor and pulsing with H_2 gas at the same temperature leads to the desorption of the remaining surface Cl as HCl. Layer-by-layer growth proceeds as long as the substrate temperature and the partial pressure of the SiH_2Cl_2 are kept low. Growth by this technique is also selective to SiO_2-masked regions. A concern with both of these approaches is that the Si surface will remain unpassivated at some stage in the sequence, providing an increased opportunity for incorporation of impurities.

In situ studies of the adsorption of $SiClH_3$ and SiH_2Cl_2 on Si(100) show that while $SiClH_3$ is not a suitable precursor for ALE growth due to substantial surface coverage of hydrogen, SiH_2Cl_2 has much lower coverage of hydrogen and can lead to single-monolayer coverage of Cl at temperatures near 500°C [25]. However, since desorption of HCl at 500°C is significant, SiH_2Cl_2 adsorption is not strictly self-limiting. As a result, other approaches have been attempted in order to achieve well-controlled layer-by-layer growth and to lower the growth temperatures. An alternate approach is that of using alternating exposures of Si_2H_6 and Si_2Cl_6 to maintain chlorine and hydrogen surface termination [26]. At 465°C, the film growth rate is roughly two monolayers per cycle (one cycle equals one Si_2H_6 and one Si_2Cl_6 exposure), and the desorption of surface hydrogen by Si_2Cl_6 dosing is a self-limiting process. Desorption of surface Cl by dosing with Si_2H_6, on the other hand, is not strictly self-limiting and is kinetically controlled in this temperature regime. Atomic hydrogen has been used to complete the Cl–H exchange reaction in a self-limiting way [27] with 400°C Si_2Cl_6 exposure. This results in Si adsorption until Cl fully terminates the surface, making the Si deposition step self-limiting. The terminating Cl layer is removed by exposure to atomic hydrogen. At 400°C, H_2 desorbs rapidly from the surface, regenerating the surface dangling bonds for the next Si_2Cl_6 adsorption. More recent work [28] has suggested that precursors such as Si_2Cl_6 and SiH_2Cl_2 are not ideal for ALE because they adsorb dissociatively, and one of the Cl atoms of these precursors fills a surface site, resulting in less than one monolayer coverage of adsorbed Si–Cl. By using SiH_2Cl_2 and increasing its residence time in the reactor, monochlorosilane—SiClH—may be produced. SiClH is a very desirable precursor for ALE growth of Si because when it adsorbs on a free Si surface, it produces H_2 and leaves a monolayer of Si–Cl, i.e., a perfectly Cl-terminated surface. Atomic hydrogen is again used to complete the cycle by desorption of HCl. Monolayer growth per cycle is achieved over a range of temperatures from 550°C to 610°C. This growth technique may not lend itself to the production of electronic-quality material, since it relies on keeping the temperature high enough to desorb hydrogen from the surface, leaving free Si sites which can lead to the incorporation of contaminants. However, it illustrates an important additional concept pertinent to ALE and ALD techniques; there are some choices of reactants, as we have seen above, which can result in self-limiting growth but not full monolayer-by-monolayer growth. Such techniques may provide deposition parameters independent of average growth rates but may also lead to local nonuniformities and roughening.

At present, a path to a technologically useful ALE growth technique for Si is less clear than for III–V compounds and other materials. What is lacking is an ideal exchange reaction wherein, for example, a hydrogen-terminated surface can be used and a Si precursor can adsorb without throwing off excess Cl that will react with some of the hydrogen-terminated sites, blocking full monolayer coverage. Additionally, most of the current options require either a source of atomic hydrogen or a photodissociative step, which may not be practical. Nonetheless, with additional progress, the ALE of Si, and by extension SiGe, could provide significant leverage for future device fabrication. Raised source/drain extensions fabricated without pattern-loading effects and better selectivity or precisely controlled channel regions of strained Si FETs are just some examples of technology elements that could be enabled by ALE of Si and SiGe.

Atomic-layer deposition of silicon nitride was demonstrated in 1996 [29] by repeated, periodic plasma nitridization of Si alternated with Si monolayer deposition from $SiCl_2H_2$. Remote plasma nitridization is carried out with NH_3 as the nitrogen source, and self-limited layer growth occurs over a range of substrate temperatures from 250°C to 400°C and for plasma powers greater than 40 W. The growth rate saturates at one-half monolayer per cycle. In more recent work, these films were used to create a stacked dielectric consisting of a 2–3-nm-thick thermal oxide followed by two monolayers of ALD SiN [30]. While most techniques for the introduction of nitrogen in the gate dielectric lead to incorporation at the lower interface, where the degradation of device performance can result, this stack has the advantage that nitrogen is incorporated away from the Si interface. Capacitors were demonstrated with good resistance to boron diffusion through the dielectric, low flat-band shift due to fixed charge, reduced tunneling currents, and good dielectric breakdown characteristics.

Other more exploratory work has pushed the atomic-layer deposition of SiO_2 down to room temperature through the use of catalyzed sequential surface reactions [31]. The atomic-layer deposition of SiO_2 has focused on breaking up the $SiCl_4 + 2H_2O \rightarrow SiO_2 + 4\ HCl$ reaction into two half-reactions,

$$(A)\ Si\text{–}OH^* + SiCl_4 \rightarrow SiO\text{–}Si\text{–}Cl_3^* + HCl \quad (6)$$

and

$$(B)\ Si\text{–}Cl^* + H_2O \rightarrow Si\text{–}OH^* + HCl, \quad (7)$$

where * represents the surface functional group. At room temperature and low pressures, the reaction rate between $SiCl_4$ and H_2O is negligible, but when the reaction is carried out in the presence of a strong Lewis base such as NH_3, catalysis of the reaction at low temperatures occurs because of the interaction of the Lewis base with the surface functional groups. The NH_3 forms a salt that complexes with the HCl reaction products in both half-reactions. Multiple pulses are used for half-reaction (A) in order to remove the reaction byproducts. A number of individual $SiCl_4$ pulses are provided in the presence of a 75-mTorr background pressure of NH_3. Half-reaction (B) is carried out with a direct mixture of H_2O and NH_3, which is allowed to react with the surface for several minutes and then removed from the chamber by means of a vacuum pump. At high enough H_2O exposure time, NH_3 pressure, and number of $SiCl_4$ pulses, the reaction saturates at one monolayer per -A-B- cycle at room temperature. The catalysis of surface reactions by gas-phase reagents represents a significant breakthrough that may be applicable to other ALD and ALE systems and may open up other materials systems to self-limiting reactions.

As discussed previously and illustrated in Figure 3, the gate dielectric is the film with the most stringent thickness and uniformity control requirements in the ITRS roadmap. Thermal oxidation of Si has served the industry well for many years, and it continues to deliver extremely good uniformity and reproducibility because of the nonlinear growth rate with time in this system. Since thermal oxidation is an almost self-limiting reaction, tolerance concerns will not be the reason for the demise of thermal SiO_2 and oxynitride as gate dielectric materials; rather, the total film-thickness requirement will approach a few monolayers as the requirement for decreasing equivalent-oxide thickness is maintained.

Figure 9 is a plot of the number of monolayers of deposited film for different choices of gate dielectric and gate conductor in order to fulfill the gate-dielectric equivalent-oxide thickness requirements as prescribed by the ITRS roadmap [1]. Bear in mind that the number of monolayers is calculated here as an effective gauge of where the material will fail to be continuous, with the full realization that most of the materials under consideration

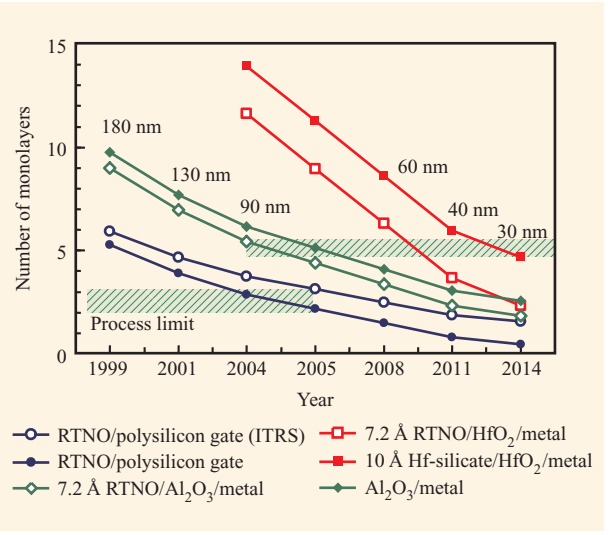


Figure 9

Equivalent number of monolayers of gate dielectric for several different gate dielectric and gate conductor combinations. The left-hand and right-hand crosshatched regions respectively indicate the practical limits for processing of single- and multiple-layer dielectric films. For dual-layer films, the thickness of the underlayer is considered fixed at a minimum value for the rapid thermal nitrided oxide (RTNO) or Hf-silicate films, respectively.

for advanced gate dielectric (and all in this plot) are actually amorphous. In addition, dielectric leakage and reliability have been ignored in this plot and are assumed not to be the limiting factors in choosing when to move to the next materials set for gate dielectrics. This oversimplification is actually not a bad assumption, given that voltage scaling has provided relief to both of those parameters and will continue to do so. Other papers in this issue will discuss reliability in more detail; nonetheless, it is instructive to consider this plot as an absolute, most optimistic view of the point at which a given material will no longer fulfill the ITRS requirements, based simply upon physical thickness.

The open circles in Figure 9 are for an oxynitride film with a dielectric constant assumed to be 1.3 times that of pure oxide and a polysilicon gate with 10% depletion of dopants as assumed by the ITRS roadmap. The filled circles are for the same materials set, but with the more realistic assumption that the gate activation will not be maintained at a constant value as assumed by the ITRS, but rather will degrade due to increased band bending at the thinner EOT. For either set of assumptions, this plot shows that polysilicon/oxynitride gates may be suitable up to the 90-nm node, but are not likely to be acceptable at the 60-nm node because films of less than two to three monolayers will be required. The case of metal on silicon

oxynitride has also been considered, although it is not plotted in the figure. The curve would fall slightly above the open circles but would probably not extend oxynitride to 60 nm.

The open diamonds in Figure 9 are for the case of Al_2O_3 as the dielectric material, with a silicon–oxynitride film as an underlayer and a metal gate electrode. Since this is a bilayer dielectric, the minimum thickness for consideration is doubled in this case to about five monolayers, two to three monolayers of each film. As a result, it is clear that this type of stack would reach practical limits at about the same time as the much simpler oxynitride/polysilicon stack. This is the most realistic assumption for Al_2O_3, given the negative results that are typically obtained when no interfacial film is used, but it indicates that if Al_2O_3 is to be used, it must be done without the benefit of an underlayer. In that case, the filled diamonds in the chart should be considered and compared against the 2–3-monolayer limit. On this basis, Al_2O_3/metal gate stacks could be useful until the 40- or even 30-nm nodes are reached, if the interface issues can be resolved.

The open squares in Figure 9 represent the case of a stack of HfO_2 on silicon–oxynitride with a metal electrode, and the filled squares, the case in which the oxynitride is replaced by a higher-dielectric-constant Hf silicate that can be formed by direct interaction of HfO_2 with the Si substrate. Both of these cases should be compared to the upper five-monolayer processing limit. The curves show that while HfO_2 on oxynitride may be suitable for the 60-nm node, continued scaling will require the use of a silicate or other higher-dielectric-constant interlayer. While there are many other issues associated with successful implementation of high-*k* gate dielectrics, such as interface states, mobility degradation, and reliability, this plot can give an indication of what the deposited film-thickness requirements will be.

Many high-*k* materials are currently under consideration, but if CVD techniques are utilized, the adequate uniformity and control currently experienced with thermally grown dielectrics cannot be taken for granted. Conventional CVD techniques are subject to either kinetically controlled growth reactions or diffusion-limited control. In either case, maintaining uniform reactant flux and temperature over large wafers will not be easy, even for the deposition of a gate dielectric for a conventional device structure, where the surface is quite planar during dielectric deposition. More advanced structures, such as replacement gates [32] or horizontal double-gated devices [33], could subject the deposition process to severe microloading. The technique of ALD is ideally suited to overcome these concerns, and gate-dielectric deposition may be the area that is furthest along in terms of implementation of ALD techniques into high-performance CMOS. As an example, consider the deposition of Al_2O_3 by ALD [34]. Monolayer control of the growth process is achieved by sequential self-limited adsorption of trimethylaluminum and H_2O at a substrate temperature of 300°C. ALD films were utilized in the fabrication of 80-nm FETs [35], and the tooling for ALD deposition of Al_2O_3 is commercially available from a number of equipment suppliers in pre-production versions.

Recently, in addition to the deposition of semiconductors and dielectrics by ALE and ALD, metals such as W [36] and metal nitrides such as TiN [37] and WN [38], which are suitable materials for contact liners, have been deposited. For the case of TiN [37], tetrakis-ethylmethylamino-titanium (TEMAT) and NH_3 are used as the sources of Ti and N, respectively. They are passed sequentially over samples heated to temperatures from 150°C to 240°C, with N_2 gas purges in between to prevent gas-phase reactions. Monolayer deposition per cycle is achieved for temperatures up to 220°C, with an increasing deposition rate at temperatures above 230°C due to the prevalence of gas-phase reactions. The films have a resistivity of <230 $\mu\Omega$-cm for deposition temperatures below 190°C, and show good performance as diffusion barriers to Cu at temperatures up to 600°C using unpatterned wafers. The ALD technique was also used to deposit films into high-aspect-ratio trenches to observe step coverage. A maximum-starting-aspect-ratio trench of 6:1 was filled to the point of >60:1 aspect ratio, with $>80\%$ bottom step coverage and no signs of pinch-off of the trench. This result and the work on other barrier and metal materials such as WN [38] and W [36] offer exciting prospects for meeting the stringent requirements future scaling will place on the deposition of metal liners and fill. Tooling for ALD contact-liner deposition is also available in pre-production form.

Another, much more exploratory area of investigation is the growth of selective ALE. Because ALE and ALD by nature are very surface-sensitive, it is possible to conceive of reactions that will proceed on some surfaces and not on others. In one example, GaAs has been deposited selectively on patterned GaAs areas opened through a Ga_2O_3 mask [39]. The condition in which there is no GaAs deposition on the masking material is a function of the hydrogen pressure in the reactor. AlGaAs, in contrast, deposits on the mask over the entire range of pressures. In another example [40], investigators took advantage of high crystallographic selectivity observed during ALE of GaAs to grow quantum wires. At high growth temperatures and with long hydrogen purge times after exposure to AsH_3, no GaAs growth on the GaAs (111)A and (110) planes is observed, while GaAs ALE growth occurs on the GaAs(100) plane. While in this example these investigators used this phenomenon to their advantage by fabricating quantum wires, it points to

the fact that ALE and ALD growth must be very well understood for each materials system in order to obtain uniform growth when that is the desired outcome.

Other concerns about ALD have been low growth rate and contamination. As the dimensions of structures are decreased, so are the thicknesses of deposited films required in their fabrication. As a result, the time will come when the deposition time required for an ALD step will be acceptable not so much because the deposition rates have been increased significantly but because the needed film thickness has been reduced. As for contamination, it is clear that many of the studies cited have focused primarily on demonstrating the feasibility of the layer-by-layer growth mode for a given materials system. With that objective in mind, investigators have found it acceptable to utilize thermal or other assisted desorption to complete one or more of the half-reactions and leave an unterminated surface. As the focus shifts toward bringing these techniques into production, deposition sequences will have to be devised that always leave the growing surface terminated with a species that is not easily removed, except by the second half-reaction. Suitable half-reactions and surface catalysts must be found in order to move ALD and ALE out of the laboratory and into the production line.

Atomic layer control in the other direction—etching

Figure 10(a) is a cross-sectional SEM image of Si trench etching for shallow-trench isolation in a 180-nm-technology node. Widely disparate trench profiles and depths are noted for features of different linewidths in this technology generation, and as ground rules are decreased, such disparities will become more pronounced. **Figure 10(b)** is a cross-sectional SEM of a contact hole adjacent to a local interconnect hole for 130-nm-technology-node CMOS. The contact hole is etched to a shallower depth because of the smaller area of the feature compared to the adjacent local interconnect of the same linewidth as the contact-hole diameter. The sidewall profile of the contact hole is also less vertical than the sidewalls of the local interconnect. Reactive ion etching (RIE) can be subject to microloading for a wide range of reasons, from the distortion of electric fields near the top of features because of mask charging [41] to limitations in the supply of reactant or byproduct due to the conductance of the trench [42] or the finite angular distribution of impinging species [43]. To alleviate these problems, it has been suggested that the only solution is the migration to lower etching pressures [44]. This may not be acceptable because of the increased charging damage observed at lower etch pressures. The challenges of microloading, anisotropy, and selectivity are sometimes in competition with one another for conventional RIE processes. Thus, other approaches may be required; one option is neutral-beam etching. A wide range of techniques have been used to generate a neutral beam of molecular and radical etchant species with translational energies from 2 to 600 eV. Current progress in this field is reviewed in this section, as well as novel atomic-layer and self-limiting etching work.

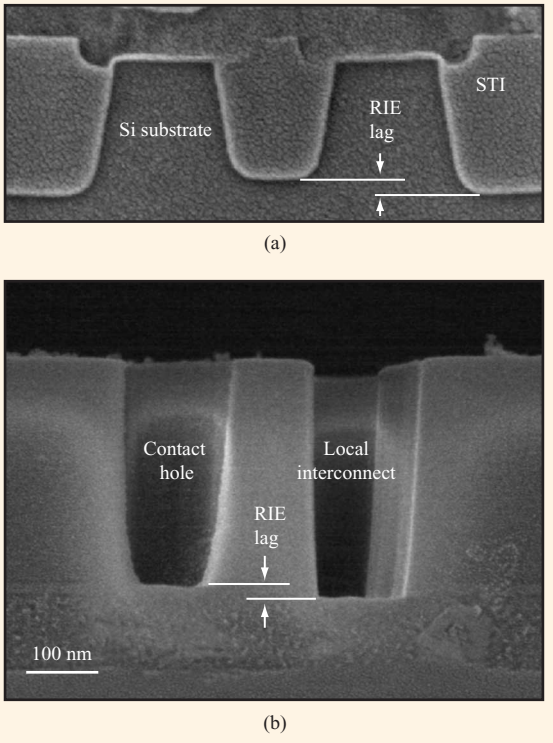


Figure 10

(a) Cross-sectional SEM image of the STI trench for 180-nm-technology-node CMOS. The RIE depth for the center trench is about 90% of the depth of the outer trenches, which have a larger lateral dimension. The sidewall profile of the center trench is also less vertical than the sidewalls of the outer trenches. (b) Cross-sectional SEM image of a contact hole adjacent to a local interconnect hole for 130-nm-technology-node CMOS. The contact hole on the left is etched to a shallower depth due to the smaller area and uni-dimensionality of the feature compared to the local interconnect on the right. The sidewall profile of the contact hole is less vertical than the sidewalls of the local interconnect.

Anisotropic etching of Si and GaAs can be performed by hyperthermal molecular [45, 46] and radical [47] beams produced by heating (and, in the case of radical production, by cracking) etchants such as Cl_2 or SF_6 in resistively heated jets that also undergo free jet expansion. The adiabatic expansion reduces the rotational temperature of the molecules and increases their translational energy. Jet temperatures from 800°C to

nearly 3000°C can be used. Early work used larger-diameter nozzles, and the beam energy was generated primarily by thermal means. As the nozzle diameter is reduced to 100 μm and the ratio of Cl_2 in He is reduced, Cl_2 molecular beams with translational energies up to 3 eV are produced with a nozzle temperature of only 910°C [48]. Beam divergence is expected to be small with this technique; therefore, anisotropic etching should be possible, but the major challenge is etching of the sidewalls by scattered species. By reducing the substrate temperature, reaction with scattered molecules is substantially reduced, and anisotropic etching by a Cl_2 hot molecular beam is achieved in the case of Si [45, 48], while Cl radical beams cannot be used because scattered Cl radicals will etch the sidewall. Both Cl_2 molecular [46] and Cl radical [47] beams produce anisotropic etching of GaAs structures because the reaction probability of Cl radicals with the GaAs is nearly unity, resulting in a very low flux of scattered radicals. By this technique, etch rates approaching that of RIE are achieved, although over beam areas of only about 1 cm^2. Scaling of this technique to full wafer-scale processing will require an array of such beams, possibly combined with wafer rastering. Contamination is also an issue with this technique because of the high source temperatures.

Higher translational energies have been obtained by the use of a pulsed laser detonation source [49, 50]. This technique utilizes a 1-mm nozzle with a 125-psig source of SF_6 pulsed into a vacuum chamber with a precisely timed CO_2 laser pulse, producing a neutral beam consisting of F and S radicals. The F translational energy can be tuned from several to 18 eV. Etching of a room-temperature Si substrate with 4.8-eV F radicals results in significant undercutting of masked regions due to the low initial reaction probability and the scattering of F radicals to the sidewalls. When the energy is raised to 18 eV, undercutting is substantially reduced, but microtrenching becomes more severe. A Monte Carlo model taking into account reaction probabilities and angular desorption distributions as well as scattering can accurately predict both the undercutting and microtrenching observed in profiles as a function of the etching conditions [50]. Etch rates by this technique are approximately 30 nm/min.

Still higher translational energies have been used for etching by neutralizing energetic ions generated by conventional plasma sources. Energies for this approach range from tens of eV to keV levels. Neutralization is accomplished in a number of different ways: by electrostatically deflecting charged species [51], through the use of multi-aperture electrodes at the beam source combined with reflection grids over the sample [52], by utilizing cusped magnetic fields to deflect ions and electrons [53], or by reflecting the beam off a surface [54]. These approaches offer various tradeoffs in terms of beam density and energy, scalability to large area, beam divergence, and contamination.

SiO_2 can be selectively etched to polysilicon by utilizing a magneto-microwave source to generate ionized Ar and CHF_3 as well as ionized radicals produced by the dissociation of CHF_3 [55]. The ions are extracted at anywhere from 400 to 600 eV, and multiple apertures are used to facilitate the transfer of the ion's charge but not its kinetic energy to thermal neutral species. In this system, etching occurs due to the kinetic energy of a neutral beam of Ar impinging on the substrate where neutral radicals are adsorbed. The neutral radicals and the energetic neutral Ar beam are generated in two separate coaxial plasmas. This is termed neutral-beam-assisted etching. By this means, anisotropic SiO_2 etch rates of more than 50 nm/min are obtained over a 200-mm wafer. Though uniformity has been limited to about 8% by the production of a uniform beam, this technique is promising in that it can be scaled to large area with reasonable etch rates and reasonably good angular distribution within the beam (estimated at about 5° in this system). However, because of the apertures and reflector grids in the path of the beam, contamination is expected to be an issue.

Other approaches have utilized a magnetic field to remove electrons and positive substrate bias to repel ions [53] and still obtain neutral-beam fluxes of 10–20-eV species capable of Si etch rates of 10 nm/min. Lower beam densities of $2 \times 10^{14}/cm^2/s$ are obtained by utilizing a "multiple bounce" technique to neutralize the beam by collision with surfaces [54]. Each such collision results in as much as 50% loss of the beam energy, but this technique has the benefit of reducing UV irradiation of the sample because of the possibility of eliminating direct-line-of-sight transport from the plasma source to the sample. It is not clear whether the angular distribution of the final beam produced by this method is adequate for anisotropic etching, and the only demonstration of this technique has been for photoresist ashing, which actually benefits from an isotropic etch. It is believed that through modification of the source, beam densities up to $1 \times 10^{18}/cm^2/s$ can be achieved, enabling resist removal at rates exceeding 1 μm/min.

Atomic-layer etching

Neutral-beam etching, as discussed in the preceding section, may hold promise for avoiding some of the current challenges faced in the RIE of structures at the advanced ground rules required for scaling, but while etchant species can be produced with sufficiently low translational energies to provide very low, controlled etch rates, etching by this technique does not proceed in an atomic-layer-by-layer fashion. In contrast, atomic-layer etching (ALET), or digital etching, was demonstrated for both GaAs and Si in 1990 [56, 57].

Etching of the (100) surface of GaAs at room temperature by ALET is achieved using a sequence consisting of a Cl_2 gas pulse, a purge cycle to remove excess gas-phase Cl_2, and bombardment with 100-eV electrons followed by another purge to remove reaction byproducts [56]. The (100) face of GaAs consists of alternating planes of either Ga or As. Two cycles are required for the removal of one GaAs layer. The etch rate is independent of the Cl_2 exposure dose, with a value of about 0.1 nm per cycle. Although this reaction is self-limiting, the etch rate of about 0.1 nm per etch cycle is lower than the value of 0.14 nm per cycle expected for single-layer removal in this materials system. Replacing the 100-eV electrons with 25-eV Ar^+ ions [58] produces a slightly different behavior. The etch rate increases with Cl_2 exposure dose up to a self-limiting value of about 0.2 nm per cycle. Above that Cl_2 dose, a longer incubation time for etching is seen because sufficient Ar^+ irradiation is needed to remove the excess adsorbed Cl_2 before ClM_x etching occurs. Though this reaction is self-limiting, the etch rate of about 0.2 nm per etch cycle is higher than expected for single-layer removal, and probably involves a component due to sputtering. In addition, the suitability of this technique for nanoscale fabrication remains unknown because the anisotropy of this type of etching has yet to be studied for GaAs.

ALET of Si can be achieved either by using Cl_2 gas adsorbed on the Si at room temperature in a manner similar to the etching of GaAs above [57] or from the cryogenic adsorption of F atoms [23]. For cryogenic etching, separate adsorption, reaction, and desorption steps are provided in a three-station reactor. First, wafers on an electrically floating sample holder are passed under a microwave discharge of $CF_4 + O_2$ which produces F. Temperatures lower than -60°C are required in order to prevent spontaneous etching. The sample is then passed under a source of UV irradiation in order to enable the reaction of F with the Si surface. Finally, the sample is passed under an Ar^+ ion beam of approximately 20 eV to desorb the reaction byproducts. The sequence is then repeated. The etch rate per cycle is self-limiting as long as sufficient Ar^+ irradiation time is allowed. However, etch rates both lower than and higher than the value for monolayer removal per cycle are obtained depending on the F-atom adsorption time, implying that the Ar^+ beam is responsible for reaction as well as desorption when excess F is available. Anisotropically etched trenches with depths of the order of 200 nm were produced by 1400 cycles of etching a patterned substrate. A substrate temperature of -160°C is required because undercutting of the mask is observed at -60°C. The sidewalls produced at -160°C are quite vertical; however, microloading was not investigated, and it remains a concern in this implementation because the etch rate is not self-limited with F exposure time but is linearly dependent.

Of additional concern with current ALET reaction sequences for GaAs and Si is the fact that processes available to date do not leave the etched surface with any terminating species at the end of each etch cycle. As a result, the surface is open to reaction with other etchants in the reactor, leading to locally more than one monolayer of etching per cycle and, in the case of Si, reaction with O_2 or H_2O in the reactor that can micromask the etch to further Si etching. Like the progression of ALE from ultrahigh-vacuum deposition systems which left no surface termination, to suitable demonstration at high pressures with hydrogen surface termination, ALET will have to migrate to chemistries which achieve better control of the surface termination. Clearly, on the basis of this criterion, ALET is in its infancy, but in principle, reactions meeting these goals should be possible to devise. Further, it should be possible to combine the concepts of ALET with neutral beam processing to achieve new, more powerful processes. For example, in all of the examples of ALET discussed above that utilize an Ar^+ ion beam or an electron beam in the process, it should be possible to replace this with Ar neutrals. Greater gains may come through the use of neutral beams to provide the correct sequence of radicals to accomplish a suitable layer-by-layer etch sequence.

In addition to ALET and neutral-beam etching, two recent processes have been developed in Si technology that are self-limiting but do not employ strictly atomic-layer-by-layer removal mechanisms. One is for the etching of Si [59] and the other for SiO_2 removal [60]. In the first example, a process consisting of sequential ozone oxidation followed by *ex situ* aqueous HF etching was developed for the purpose of ultrashallow depth profiling with X-ray photon and Auger electron spectroscopy. The Si removal per cycle is determined by the very reproducible and somewhat self-limiting oxidation of the Si surface by ozone exposure. Removal of 0.5 nm per cycle was achieved, indicating that more than a monolayer removal of the surface is obtained. Since this work was carried out for the purpose of depth profiling, total film removals were limited to <10 nm in the study, and it is not clear whether roughening of the surface will occur after many cycles. Also, since wet etching is used, one would not expect this process to be anisotropic, which may limit its usefulness. In the second example [60], a self-limiting reaction for removing SiO_2 was reported. HF and NH_3 are reacted with oxide at room temperature in vacuum to form ammonium hexafluorosilicate, $(NH_4)_2SiF_6$, which remains on the surface and serves as a diffusion barrier to further reaction of the HF with the SiO_2 surface. After the reaction is stopped, the reaction byproduct is removed by thermal desorption at >100°C or by dissolution in H_2O. This results in the reaction being

self-limited to approximately 12-nm removal of SiO_2. If removal of thicker films is desired, repeated sequencing can be carried out, or the reaction can be carried out at higher temperatures to produce a thinner byproduct layer. Another very interesting feature of this reaction sequence comes from the fact that the byproduct undergoes a volume expansion of a factor of 3 with respect to the SiO_2 film removed. The byproduct will fill in cracks, scratches, or grooves in SiO_2 during etching [61], thereby locally reducing the reaction rate and thus reducing surface roughness.

Applications of ALE, ALD, and ALET

ALE, ALD, and ALET techniques will be required in order to continue technology scaling as mandated by the ITRS roadmap. This is true for continued scaling of conventional device structures as well as for new device structures such as replacement gate or damascene gate structures, devices fabricated with high-*k* gate dielectrics, FinFETs [62], and other novel devices. Conventional device scaling will require film thickness values and tolerances as plotted in Figures 3 and 4, and will be subject to the RIE lag and high-aspect-ratio fill issues illustrated in Figures 5 and 10. In addition, precise control of the gate electrode dimension and profile across a wide range of pattern factors will be necessary. Thus, for the fabrication of conventionally scaled CMOS devices, many of the RIE and CVD process steps could benefit from ALET and ALD processes. However, even if structures with the correct physical dimensions and tolerances can be fabricated, dopant activation and junction depth and abruptness face fundamental limitations when fabrication is done by means of implantation and thermal activation, as indicated in Figures 7 and 8. As a result, continued scaling of conventional device structures will also require raised source/drain junctions with a level of control not attainable through CVD techniques, necessitating raised source/drain depositions by selective ALE.

Looking specifically at selective ALE in raised source/drain applications, the growth of *in-situ*-doped films could provide a path for additional scaling of R_{ext} by improving ρ_e, $\rho_\square$, R_{ac}, and R_{sp}. Epitaxial growth of *in-situ*-doped films at low temperatures by CVD has already demonstrated very abrupt junction profiles because of the lack of implant damage combined with very low thermal cycles and incorporation of active dopants well above the solid solubility limit due to their incorporation on substitutional sites during growth [63]. In addition, junction abruptness is decoupled from the junction depth when a recess is combined with selective growth, as in the buried-source/drain [64] or recessed-junction approaches [12]. These approaches require a recess of the extension-junction region, followed by selective deposition of the extension junction. For the recess to be practical, it will be necessary to achieve uniformity over large-area wafers with insensitivity to pattern factor. By utilizing self-limiting processes with atomic-layer control of etching, this should be achievable. Epitaxial growth uniformity will also be improved through the use of ALE techniques, but additional device benefits may be gained through the use of strain-modulated epitaxy to incorporate even higher levels of active dopants. For the case of p-FETs, SiGe growth provides compressive stress which favors the incorporation of the smaller B atom on substitutional sites [12]. By this technique, active doping levels of more than 5×10^{20} cm^{-3} are obtained. Active doping concentrations of 1×10^{20} cm^{-3} and 2×10^{20} cm^{-3} are achieved for P- and As-doped Si at low temperatures [63]. The chemical level of the dopants is two to five times greater than the active concentration, so hyperactivation of n-type impurities may also be achievable for n-FETs through the use of local strain engineering.

In addition to improving the R_{ext} by modifying the structure of the extension and the placement of dopants, Equation (4) indicates that the R_{co} component of R_{ext} can be reduced by reduction of ϕ_b. For a single silicide, reduction of the barrier for the n-FET contact will degrade the p-FET, or vice versa. Replacement of the extension junction with dual Schottky source/drains has been modeled [65] and fabricated [66]. Materials that provide barriers of less than 0.1 V for both n and p will probably be required. Without an extension junction, the thermionic emission current at low gate voltages may be insufficient to support the channel current. As a result, some extension doping is also likely to be required. For the case of low interface doping concentrations, 100 times reduction in contact resistance has been demonstrated by using embedded nanocrystals to significantly enhance the tunneling current [14]. In that work, nanocrystals were formed on Si by evaporating a thin layer of a material that is more likely to ball up and produce nanocrystals to reduce surface free energy than to form a uniform thin-film silicide upon annealing. In this case gold was used, but other materials could be used as well. A contact material with a dissimilar work function was then deposited over the Si/nanocrystal surface. At triple points between the contact metal, Si and nanocrystal, the electric field was enhanced, thus increasing the field-emission current. However, this is applicable predominantly for Schottky contacts to lightly doped semiconductor, where the field enhances the total tunneling appreciably. Contact resistances as low as 20 to 100 $\mu\Omega$-cm^2 were obtained. As scaling of SOI CMOS devices pushes body thickness to less than 30 nm, the specific contact resistivities obtained with nanocrystals are from 1×10^3 to 1×10^4 times higher than what would be required to keep the R_{ext} below 100 Ω-μm, if R_{ext} is purely limited by R_{co} with a contact area determined by T_{Si}, the silicon film thickness. Thus,

the example of utilizing nanocrystals to modify the interface is used only as an example of the type of interface engineering that may be required for future devices.

If continued scaling of the conventional device structure is not sustainable, alternate device structures such as replacement gate structures [32] or FinFETs [62] may be required in order to maintain the trend of improved device performance every generation. The replacement gate approach may be required for the successful fabrication of high-*k* gate dielectric devices, either with polysilicon or metal gates, because this process flow permits device fabrication with the high-temperature steps for junction formation preceding the gate dielectric. This sequence reduces the thermal processing of the high-*k* material which is crucial for the integrity of the interface. In this sequence, devices are processed with sacrificial gate electrodes, which are removed later in the process, allowing the dielectric deposition and gate electrode fabrication in troughs left by the sacrificial gate. Deposition of uniform dielectric films and sidewall spacers in troughs with variable pattern factors and widths will have to meet tolerance requirements at least as stringent as for the conventional device as quantified in the ITRS. Uniformly recessing the channel region before gate dielectric deposition is an additional desirable step that can result in increased device performance. However, deposition and RIE inside troughs will be subject to more mass-transport issues than the same processes carried out for conventional structures, making these applications well suited to ALD and ALET.

The FinFET is another alternate device structure with numerous potential applications of atomic-layer techniques in its fabrication. In order to maintain control of short-channel effect as channel length is scaled below 20 nm, double-gated device structures have been proposed [62]. Channel lengths as short as 10 nm should be attainable with manageable off currents, if the body thickness can be scaled to less than 5 nm. Fabrication of the thin body of a FinFET has typically been carried out by e-beam lithography and etching in laboratory experiments, but in order to lead to improved SCE control, the T_{body} is required to be $\lesssim 1/4\ L_{gate}$. Assuming that the absolute limiting lithography is used on the gate definition, body patterning will be by definition sublithographic, and a nonlithographic technique will be required. Sidewall image transfer is one possible technique for body formation. In this scheme, a sacrificial material is first deposited on an SOI substrate from which the Fin is fashioned. A vertical step is patterned into the sacrificial material by lithography and RIE selectively stopping on the SOI. A conformal spacer is then deposited and etched, leaving sidewalls of material that will be used as a hard mask for definition of the Fin once the sacrificial material is removed. A variation of just one degree in the angle of the sacrificial material onto which the spacer is placed will lead to about 1 nm in Fin-body-thickness variation for a Fin height of 50 nm. Calculations show that control of the Fin-body thickness to within 10% is required for control of the threshold voltage to within ±30 mV [32]. ALET will be required in the definition of the step in the sacrificial material to achieve the required control over sidewall angle across all pattern factors. The conformal spacer will require ALD in order to ensure accurate thickness control over all pattern factors. The sidewall image transfer process will require ALET in order to achieve the necessary anisotropy and pattern factor insensitivity. The sidewall spacers and raised source/drain will require ALD and ALE to achieve suitable film-thickness tolerances.

Substrate engineering—the road to new materials

A move away from SiO_2-based dielectrics as dictated by gate dielectric scaling is a move away from one of the fundamental advantages of Si-based CMOS over other materials systems—the Si/SiO_2 interface. That may warrant a re-evaluation of the materials set for future devices. The continuation of current scaling trends, as set forth in the ITRS roadmap, will probably require the semiconductor industry to develop a new gate dielectric material within the next several years; assuming that that challenge can be met, physical gate lengths in the vicinity of 20 nm will be in production by the middle of the next decade. Other papers in this special issue discuss why controlling the short-channel effect at dimensions smaller than these could limit further scaling of Si-based devices and thus motivate the need for new materials for device fabrication.

For a comparison of some mechanical and electrical properties of a selected group of semiconductor materials, refer to **Table 1**.

Because the mobilities of materials other than Si are attractive, consider some of the alternatives to Si beyond the SiO_2 era. The industry will be faced with a choice: direct fabrication of substrates having diameters of 300 mm or greater for a new materials set, or epitaxial growth of a new material on Si substrates. The requirement for a large wafer size will be set by the level of integration already achievable in Si technology. Direct growth of other semiconductors from the melt has been carried out, but wafer sizes are limited, and material quality is poor because of the loss of column V elements significantly below the melting point for the III–V compounds. As indicated in Table 1, none of these materials has a hardness value as high as that of Si, so handling them will be more difficult. The most likely path for the introduction of new materials will be to utilize the Si

Table 1 Mechanical and electrical properties of some semiconductor materials.

	Si	*Ge*	*GaAs*	*InP*
T_m (°C)	1415	937	1238	1062
Lattice constant (A)	5.4309	5.6461	5.6532	5.8687
Thermal expansion coefficient (10^{-6}/K)	2.5	6.1	5.4	4.6
Microhardness (N/mm^2)	11270	7644	7500	4100
Mobility at 25°C (cm^2/V-s)	n – 1500	n – 3800	n – 8800	n – 4600
	p – 450	p – 1820	p – 400	p – 150
E_g (eV)	1.1	0.67	1.35	1.27

substrate and all of the infrastructure that it enables, but heteroepitaxy of these materials will be difficult by conventional means because of the large lattice mismatch and the differences in thermal expansion coefficients.

One path might be the homoepitaxial growth of active layers of these materials on wafers of the right size, but with less than adequate perfection, on the assumption that at epitaxial growth temperatures (which are significantly lower than the melt), the equilibrium concentration of vacancies is greatly reduced and epitaxial layer quality may be adequate. The film might then be transferred to a handler wafer of another material such as Si, using bonding and etchback. Point defects would be reduced by this means; however, extended defects would be transferred into the active device layer. More recent developments in heteroepitaxy have opened up an alternate path. These developments are discussed in the next section, after a brief review of the traditional limits to heteroepitaxial growth of mismatched films due to strain relaxation.

For heteroepitaxial growth of a film that does not have the same lattice constant as the substrate on which it is growing, the epitaxial layer can grow with perfect crystalline alignment to the substrate, but the grown crystal lattice will be tetragonally distorted. Elastic strain will not result in the generation of dislocations until a critical layer thickness is exceeded. There are two well-known models for the prediction of critical layer thickness [67, 68]. One considers the bending force required to bend threading dislocations in the plane of the interface in order to relieve misfit in the system, and can be thought of as an equilibrium model. The other considers the self-energy of creation of a dislocation. In both models, the force or energy is provided by the mismatch strain in the system, which is proportional to the natural log of the layer thickness. Because the energy for creation of dislocations is greater than that required to move pre-existing threading dislocations, the resulting critical layer thickness is about one order of magnitude greater than would be predicted by the dislocation movement model. Films with a thickness above the dislocation motion model but below the dislocation creation model predictions can be thought of as metastable, because any imperfection can lead to the relaxation of the layer. The above treatments assume that the substrate is, in effect, of infinite thickness. Ideal misfit dislocations are pure edge dislocations at the growth interface; however, in both Si and GaAs systems, 60° dislocations are common and result in both the relief of mismatch and a threading component into the epitaxial film. One can estimate the aerial density of misfit dislocations, N_m, required to fully relax a heteroepitaxial film on the basis of the relaxed lattice constants of the films being considered, a_1 and a_2 [69]:

$$N_m \approx 4/(1/a_1^2 - 1/a_2^2), \tag{8}$$

which gives about 2×10^{12} dislocations/cm^2 for a mismatch of 1×10^{-3}. For systems with a few percent mismatch, such as SiGe on Si, relaxed buffer layers have been grown with dislocation densities below 1×10^5/cm^2 by growing graded buffer layers in an attempt to "steer" the threading segments out toward the wafer edges or annihilate one another. Achieving these dislocation densities, which are still too high for most device applications, requires film thicknesses greater than a micron.

In 1991 Lo [70] suggested that it should be possible to grow a pseudomorphic structure of any thickness if it is grown on a freestanding substrate that is thinner than the critical thickness for the mismatch in the system. In a sense, this hypothesis turns conventional heteroepitaxy on its head by treating the epitaxial layer as the substrate and the substrate as the film to which the critical thickness applies.

The result is an effective critical thickness (t_{eff}) that is related to the conventional critical thickness (t_c) but modified by the thickness of the substrate (t_s) as follows:

$$1/t_{eff} = 1/t_c - 1/t_s\,. \tag{9}$$

For $t_s < t_c$, the effective critical thickness is infinite, and even for cases in which t_s approaches t_c, the effective critical layer thickness, t_{eff}, becomes quite large.

The first demonstration of this concept was the growth of a 200-nm-thick InGaAs layer on freestanding 80-nm-thick GaAs membranes [71, 72]. This growth thickness is about double the conventional critical layer thickness. X-ray diffraction showed that, unlike samples grown on conventional GaAs substrates, the samples grown on the membranes were not relaxed and did not have misfit dislocations in the epitaxial film. Cross-sectional TEM

revealed that dislocations were pulled down into the substrate film, illustrating a second important concept predicted by the theory [70]. For substrates thinner than the epitaxial film, the image force pulls dislocations into the substrate, introducing a very effective dislocation-gettering mechanism. Thus, even for the case in which $t_s > t_c$, this mechanism will allow the growth of layers exceeding the critical layer thickness, but without misfit dislocations threading into the epitaxial film.

While the work on freestanding membranes demonstrates the concept and may be useful for certain niche applications, it is not a practical solution for gigascale integration. Work that has followed extends the concept in an effort to achieve a "virtual" freestanding substrate on commercially available and useful physical substrates. This has come to be known as a "compliant substrate." Numerous approaches have been tried, but they can generally be categorized into two classes:

1. Placing the thin freestanding substrate on an amorphous interlayer on top of a thick mechanical handling wafer.
2. Bonding the thin freestanding substrate with a rotational offset ("twist-bonded") to a thick mechanical handling wafer of the same material, such that an array of screw dislocations is formed at the bonding interface.

In 1994 Powell et al. [73, 74] examined the relaxation of layers of $Si_{0.85}Ge_{0.15}$ grown on thin SOI that exceeded the classical equilibrium critical layer thickness by a significant amount, but not the metastable critical layer thickness. A SOI film was first thinned to 50 nm, and 180 nm of SiGe was grown at 500°C. The layers were subsequently annealed at temperatures from 700°C to 1050°C. X-ray diffraction showed that the layers are fully strained after growth, with no dislocations detected, but are partially relaxed by annealing at 700°C. TEM analysis showed that the misfit dislocations that are present at the Si–SiGe interface after annealing thread down into the Si substrate rather than up into the epitaxial layer, with a density exceeding $1 \times 10^7/cm^2$, while no dislocations within the detection limit are detected in the SiGe. Further annealing at 900°C has no effect on the strain or dislocations, indicating that the film is relaxed to the equilibrium condition, where the strain can no longer push threading dislocations. In this case the growth and annealing temperatures were too low to allow for the tensile strain in the SOI to be accommodated by viscous reflow of the buried oxide, so it was postulated that plastic deformation at the Si/SiO_2 interface took place. It should be noted that the epitaxial layer deposition took place through a shadow mask in this study, so growth areas were limited.

In later work, 1 μm of $Si_{0.6}Ge_{0.4}$ was grown at 500°C on thin SOI and shown to be "relaxed" as grown, without requiring post-growth annealing [75]. This was confirmed with TEM only; no X-ray diffraction was done. The TEMs show a dense array of dislocations in the SOI film with none threading into the SiGe layer, whereas in a control sample of the same thickness and composition grown on bulk Si, $>2 \times 10^{11}$ dislocations/cm^2 in the SiGe are seen. What is clear from this study is that for mismatched growth on thin SOI, the dislocations are directed downward toward the substrate by the image force, leading to misfit accommodation even at low growth temperature. The degree of relaxation for this example is unknown.

Other workers have added boron to the buried oxide of SOI wafers in order to achieve viscous reflow at growth temperatures [76, 77]. SiGe epitaxial films on such SOI substrates were from 38% to 64% relaxed. Relaxation was a function of boron content in the buried oxide (BOX) in one of the studies [77]. Further relaxation, up to 95%, can be obtained by post-growth annealing at 1000°C [76]. Most of the relaxation is accommodated by threading of misfit dislocations downward into the substrate. Misfit dislocation densities as low as 1×10^3 cm^{-2} were observed in partially relaxed films. This approach may not be useful for MOS device applications with such a large source of boron in the vicinity of the channel, but it demonstrates that strain in the epitaxial film can be relaxed beyond the lower limit for motion of misfit dislocations based upon the equilibrium model. This result suggests that there are at least two classes of "compliant" substrates; the first [73, 74] relieves the mismatch strain in the substrate utilizing the motion of dislocations. Since finite strain is required for dislocation glide, complete relaxation cannot occur. The second [76, 77] approach also uses the motion of dislocations to relieve strain, but additionally, at temperatures above that for viscous flow, it provides a mechanism for nearly complete relaxation because dislocation motion in the crystal is not required. Other workers have used viscous bonding interfaces, particularly in the GaAs system, where an In/Ga bond and a glass diffusion barrier are used to bond a thin compliant GaAs substrate to a GaAs handling wafer [78]. In other work utilizing thin (10-nm) GaAs substrates bonded to a handler wafer with thin borosilicate glass, growth of strained $Ga_{0.91}In_{0.09}As$ films to thicknesses five times greater than the critical layer thickness was carried out by organometallic CVD (OMCVD) at 700°C. In this case, while there is a dramatic reduction in misfit dislocations threading up into the epitaxial layer when compared to growth on bulk GaAs, very little strain relaxation occurs. However, these films were not subjected to post-growth anneals, and for the short (~2 min) growth times studied, viscous flow may not have had time to occur.

SiGe layers have also been grown on porous silicon substrates [79]. Substrates consisting of a two-layer porous structure with higher porosity at the bottom followed by a lower-porosity film directly beneath an epitaxially grown thin Si buffer layer are used. $Si_{0.8}$–$Ge_{0.2}$ films grown on these substrates do not have the regular array of misfit dislocations shown by growth on bulk Si substrates.

In 1997 Ejeckam and co-workers reported that $In_{0.35}Ga_{0.65}P$ could be grown to a thickness exceeding the classical critical layer thickness on GaAs substrates by a factor of 30 [80]. This is achieved by utilizing a GaAs substrate of 10-nm thickness bonded to a thick GaAs bulk substrate with a $>10°$ twist angle between the $\langle 110 \rangle$ directions while keeping the surface normals parallel. This twist angle introduces a dense square array of screw dislocations confined to the substrate with a spacing (d) given by

$$d = |b|/2\sin(\theta/2) \approx |b|/\theta \qquad (10)$$

for small θ, where (b) is the Burgers vector and θ is the twist angle. This results in less than 2-nm dislocation spacings in GaAs for twist angles of $>10°$. OMCVD of 300-nm InGaP films at 640°C on $\theta = 17°$ twist-bonded substrates and bulk substrates was compared by XTEM examination. Growth on bulk substrates shows threading dislocations and stacking faults, while growth on the twist-bonded substrates shows no dislocations, despite exceeding the critical layer thickness by more than 30 times.

The same group later showed the growth of InSb with a 14.7% lattice mismatch on similar substrates [81]. The twist angle was increased to 40°, and the thickness of the bonded substrate was reduced to <2 nm in order to accommodate the increased strain. A 650-nm film of InSb was grown by MBE simultaneously on bulk GaAs and twist-bonded substrates. XTEM analysis shows dislocation densities exceeding 1×10^{11} cm^{-2} on bulk substrates, whereas no dislocations are present for growth on twist-bonded substrates.

Particularly exciting is the growth of GaN on Si substrates employing these techniques [82, 83]. While this work has been motivated primarily by the desire to fabricate electro-optical devices, further progress with other III–V compounds could enable high-performance, dense logic in the event that Si, strained Si, and Ge options are exhausted.

Even more exploratory work utilizes bottom-patterned compliant substrates to produce strain-modulated epitaxy [84]. In this work, epitaxial growth on the flat top surface of a bonded substrate can result in lateral material variations in the grown film. This is achieved by patterning of the substrate prior to bonding, resulting in nonuniform strain in the top surface of the bonded region once it is thinned. The resulting epitaxial layer growth is modified by the strain field by strain-dependent growth kinetics. It is possible that hyperactivation of dopants might be achieved in films grown under these conditions.

While it is recognized that a significant reduction in the misfit dislocation density threading into heteroepitaxially grown films has been achieved for a wide range of materials systems, the mechanism for long-range accommodation of the mismatch into the substrate layer is not well understood, and there is considerable dispute about it [85, 86]. Much of the work to date has been limited to TEM observations confirming limited misfit dislocation densities, but full characterization of the degree of relaxation in the heteroepitaxially grown layers is frequently lacking. While some investigators have claimed that fully relaxed films can be grown by "free slipping" at the substrate interface [73], others feel that it is highly unlikely that this can take place on a macroscopic level [85]. It is likely that in some of the systems discussed, significant strain is still present in the heteroepitaxial film, and the relaxation that has occurred is accommodated by misfits into the substrate.

In spite of the uncertainties in the physical mechanisms at this time, it is clear that these new techniques provide a significant reduction in the generation and propagation of dislocations in heteroepitaxial systems. Further, they may provide a pathway for production of superlattices and artificially engineered strain conditions, thereby modifying the electronic transport and resulting in new "designer" materials for high-performance ultralarge-scale integration.

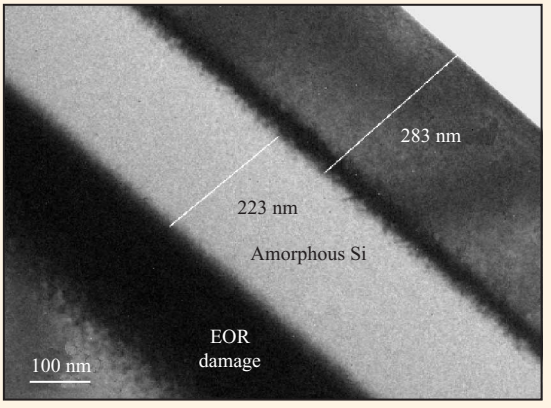
It is worth questioning the viability of such a path for the fabrication of ultralarge-scale integrated circuits. Early in the evolution of solid-state devices, many questioned the economy of building transistors in single-crystal material when a large infrastructure for polycrystalline substrates already existed because of the rectifier industry [3]. Ultimately the material that produced devices with the best characteristics was chosen, and costs and defects were driven down. SOI provides another illustration of the concept. The TEM images in **Figures 11(a)** and **11(b)** show a SOI wafer fabricated by oxygen implantation and annealing (SIMOX), before and after the annealing step. The large number of defects that are introduced by high-dose oxygen implantation are almost completely removed by the annealing step, with none present in the field of view of the TEM image. The defect density is not zero; however, the few defects that remain are benign with respect to device performance and yield. For the highest-performance CMOS devices, the improvement in device performance outweighs the incrementally small additional substrate cost, and high-performance, high-reliability SOI CMOS chips are entering their third generation in production this year. Finally, consider the practice of chemical–mechanical polishing (CMP), now prevalent throughout the FEOL and BEOL process steps. Even

the industry visionary Gordon Moore recalled that he was skeptical at first as to the applicability of CMP to semiconductor manufacturing[1] and commented that at the time of introduction he thought it to be an "unnatural act." The heteroepitaxial processes and substrate engineering discussed above require acts that are no less unnatural; they just require some atomic-level finesse. The ultimate arbiter will be device physics. If device physics dictates that the short-channel effect for devices in Si limits further scaling, alternate substrates with a suitable defect density will be engineered.

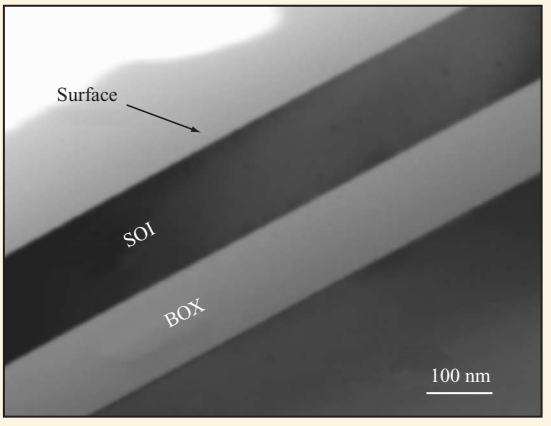
Conclusion

This paper has explored some of the process options that will enable continued scaling of Si-based CMOS over the next decade and some of the work in substrate engineering and heteroepitaxy that might provide a progression to other materials systems while maintaining the possibility of ultralarge-scale integration levels currently achievable. The need for atomic-level control of depositions and etches was shown and some possible solutions were presented, with most of the techniques discussed having been developed within the past decade. Overall, it is safe to say that manipulation at the atomic scale will be required to obtain the necessary performance and density in the Si materials system, consistent with the ITRS roadmap. ALE and ALET are some of the techniques that may be used for this purpose. The current state of the art of ALE of Si is still very immature, and that of ALET or neutral beam etching is even less mature. It is likely that these new processing techniques will be needed, regardless of whether conventional Si CMOS scaling is pursued or alternate device structures such as the FinFET or replacement gate are chosen, given the tolerance requirements and difficult fabrication sequences for most of the alternate structures that have been proposed. Additional leverage will be obtained by engineering artificial materials, either by strain, superlattices, or (as in the case of contacts) embedded nanocrystals to reduce contact resistance. These processes also are in their infancy, but they could provide some relief to the inherent physical and electrical properties of the Si–SiO_2 system at the core of our current technologies. Beyond Si CMOS it is possible that heteroepitaxy on "compliant" substrates will play a significant role in opening the door to new materials. The challenges in front-end-of-line processing are significant; however, there are important opportunities for new atomic-level manipulation techniques, which if met could enable continued performance and density scaling.

[1] G. Moore, "Silicon Device Fabrication Technology Revisited," IEDM 1996 Conference Luncheon Plenary Talk, from the private notes of Paul Agnello in attendance, 1996.



(a)



(b)

Figure 11

(a) XTEM image of an SOI wafer in the process of being fabricated. Bulk Si substrate received a high-dose oxygen implant, resulting in an amorphous region and end-of-range damage. Crystalline defects above the amorphous region are observed. (b) XTEM image of a completed SOI substrate, formed by annealing a wafer processed as in part (a) under optimal conditions to coalesce the oxygen in the substrate into a buried SiO_2 layer (BOX) and repair the crystalline damage.

Acknowledgments

I would like to acknowledge many fruitful discussions with my colleagues, Steven Bedell, Byoung Hun Lee, Subramaniam Iyer, Serge Biesemans, Devendra Sadana, and Paul Muller, and also the labors and insight of my whole development team, through whom I see the challenges of scaling CMOS technologies, thereby shaping the topics I have chosen to discuss in this paper.

References

1. *International Technology Roadmap for Semiconductors* (ITRS), 2000 update, International Sematech, Austin, TX;

see *http://public.itrs.net/*.
2. W. Shockley, "The Path to the Conception of the Junction Transistor," *IEEE Trans. Electron Devices* **ED-23,** 597–620 (1976).
3. G. Teal, "Single Crystals of Germanium and Silicon—Basic to the Transistor and Integrated Circuit," *IEEE Trans. Electron Devices* **ED-23,** 621–639 (1976).
4. M. M. Atalla, E. Tannenbaum, and E. J. Scheibner, "Stabilization of Si Surfaces by Thermally Grown Oxides," *Bell Syst. Tech. J.* **38,** 749 (1959).
5. D. Kahng and M. M. Atalla, "Silicon–Silicon Dioxide Field Induced Surface Devices," presented at the IRE–AIEE Solid-State Device Research Conference, Carnegie Institute of Technology, Pittsburgh, 1960.
6. F. M. Wanlass and C. T. Sah, "Nanowatt Logic Using Field-Effect Metal-Oxide Semiconductor Triodes," *IEEE International Solid-State Circuits Conference*, *Digest of Technical Papers*, 1963, pp. 32–33.
7. J. C. Sarace, R. E. Kerwin, D. L. Klein, and R. Edwards, "Metal–Nitride–Oxide–Silicon Field Effect Transistors with Self-Aligned Gate," *J. Solid-State Electron.* **11,** 653–660 (1968).
8. J. Birnbaum and R. S. Williams, "Physics and the Information Revolution," *Phys. Today*, pp. 38–42 (2000).
9. V. Zhirnov, D. Herr, J. Hutchby, and R. Calvin, "Future Trends in Nanofabrication and Semiconductor Research Needs in the Basic Physical Sciences," *Extended Abstracts of the Second International Conference on Microelectronics and Interfaces*, American Vacuum Society, paper J1J2MoA1, 2001, pp. 23–26.
10. M. K. Gobbert, T. P. Merchant, L. J. Borucki, and T. S. Cale, "A Multiscale Simulator for Low Pressure Chemical Vapor Deposition," *J. Electrochem. Soc.* **144,** 3945–3951 (1997).
11. K. K. Ng and W. T. Lynch, "The Impact of Intrinsic Series Resistance on MOSFET Scaling," *IEEE Trans. Electron Devices* **ED-34,** 503–511 (1987).
12. S. Gannavaram, N. Pesovic, and M. C. Ozturk, "Low Temperature (<800°C) Recessed Junctions Selective Silicon–Germanium Source/Drain Technology for Sub-70nm CMOS," *IEDM Tech. Digest*, pp. 437–440 (2000).
13. R. Kasnavi, P. B. Griffin, and J. D. Plummer, "Ultra Low Energy Arsenic Implant Limits on Sheet Resistance and Junction Depth," *Symposium on VLSI Technology*, *Digest of Technical Papers*, 2000, pp. 112–113.
14. V. Narayanan, Z. Liu, Y.-M. N. Shen, M. Kim, and E. C. Kan, "Reduction of Metal–Semiconductor Contact Resistance by Embedded Nanocrystals," *IEDM Tech. Digest*, pp. 87–90 (2000).
15. C. M. Osburn, J. Y. Tsai, and J. Sun, "Metal Silicides: Active Elements of ULSI Contacts," *J. Electron. Mater.* **11,** 1725–1739 (1996).
16. T. Suntola and J. Antson, "A Method for Producing Compound Thin Films," U.S. Patent 4,058,430, filed 1975, issued 1977.
17. M. Ahonen, M. Pessa, and T. Suntola, "A Study of ZnTe Films Grown on Glass Substrates Using an Atomic Layer Evaporation Method," *Thin Solid Films* **65,** 301–307 (1980).
18. M. Pessa, R. Makela, and T. Suntola, "Characterization of Surface Exchange Reactions Used to Grow Compound Films," *Appl. Phys. Lett.* **38,** 130–131 (1981).
19. J. Nishizawa, H. Abe, and T. Kurabayashi, "Molecular Layer Epitaxy," *J. Electrochem. Soc.* **132,** 1197–1200 (1985).
20. S. M. Bedair, M. A. Tischler, T. Katsuyama, and N. A. El-Masry, "Atomic Layer Epitaxy of III–V Binary Compounds," *Appl. Phys. Lett.* **47,** 51–53 (1985).
21. E. Colas, R. Bhat, B. J. Skromme, and G. C. Nihous, "Atomic Layer Epitaxy of Device Quality GaAs," *Appl. Phys. Lett.* **55,** 2769–2771 (1989).
22. T. Tanaka, T. Fukuda, Y. Nagasawa, S. Miyazaki, and M. Hirose, "Atomic Layer Growth of Silicon by Excimer Laser Induced Cryogenic Chemical Vapor Deposition," *Appl. Phys. Lett.* **56,** 1445–1447 (1990).
23. Y. Horiike, T. Tanaka, M. Nakano, S. Iseda, H. Sakaue, A. Nagata, H. Shindo, S. Miyazaki, and M. Hirose, "Digital Chemical Vapor Deposition and Etching Technologies for Semiconductor Processing," *J. Vac. Sci. Technol. A* **8,** 1844–1850 (1990).
24. J. Nishizawa, K. Aoki, S. Suzuki, and K. Kikuchi, "Silicon Molecular Layer Epitaxy," *J. Electrochem. Soc.* **137,** 1898–1904 (1990).
25. D. D. Koleske, S. M. Gates, and D. B. Beach, "Precursors for Si Atomic Layer Epitaxy: Real Time Adsorption Studies on Si(100)," *Appl. Phys. Lett.* **61,** 1802–1804 (1992).
26. S. M. Gates, D. D. Koleske, J. R. Heath, and M. Copel, "Epitaxial Si Films on Ge(100) Grown Via H/Cl Exchange," *Appl. Phys. Lett.* **62,** 510–512 (1993).
27. D. D. Koleske and S. M. Gates, "Atomic Layer Epitaxy of Si on Ge(100) Using Si_2Cl_6 and Atomic Hydrogen," *Appl. Phys. Lett.* **64,** 884–886 (1994).
28. E. Hasunuma, S. Sugahara, S. Hoshino, S. Imai, K. Ikeda, and M. Matsumura, "Gas-Phase-Reaction-Controlled Atomic-Layer-Epitaxy of Silicon," *J. Vac. Sci. Technol. A* **16,** 679–684 (1998).
29. H. Goto, K. Shibahara, and S. Yokoyama, "Atomic Layer Controlled Deposition of Silicon Nitride with Self-Limiting Mechanism," *Appl. Phys. Lett.* **68,** 3257–3259 (1996).
30. A. Nakajima, T. Yoshimoto, T. Kidera, K. Obata, S. Yokoyama, H. Sunami, and M. Hirose, "Atomic-Layer-Deposited Silicon-Nitride/SiO_2 Stacked Gate Dielectrics for Highly Reliable p-Metal-Oxide-Semiconductor Field-Effect Transistors," *Appl. Phys. Lett.* **77,** 2855–2857 (2000).
31. J. W. Klaus and S. M. George, "Atomic Layer Deposition of SiO_2 at Room Temperature Using NH_3-Catalyzed Sequential Surface Reactions," *Surf. Sci.* **447,** 81–90 (2000).
32. A. Chatterjee, R. A. Chapman, G. Dixit, J. Kuehne, S. Hattangady, H. Yang, G. A. Brown, R. Aggarwal, U. Erdogan, Q. He, M. Hanratty, D. Rogers, S. Murtaza, S. J. Fang, R. Kraft, A. L. P. Rotondaro, J. C. Hu, M. Terry, W. Lee, C. Fernando, A. Konecni, G. Wells, D. Frystak, C. Bowen, M. Rodder, and I.-C. Chen, "Sub-100nm Gate Length Metal Gate NMOS Transistors Fabricated by Replacement Gate Process," *IEDM Tech. Digest*, pp. 821–824 (1997).
33. G. W. Neudeck, T.-C. Su, and J. P. Denton, "Novel Silicon Epitaxy for Advanced MOSFET Devices," *IEDM Tech. Digest*, pp. 169–172 (1997).
34. E. P. Gusev, M. Copel, E. Cartier, I. J. R. Baumvol, C. Krug, and M. A. Gribelyuk, "High-Resolution Depth Profiling in Ultrathin Al_2O_3 Films on Si," *Appl. Phys. Lett.* **76,** 176–178 (2000).
35. D. A. Buchanan, E. P. Gusev, E. Cartier, H. Okorn-Schmidt, K. Rim, M. A. Gribelyuk, A. Mocuta, A. Ajmera, M. Copel, S. Guha, N. Bojarczuk, A. Callegari, C. D'Emic, P. Kozlowski, K. Chan, R. J. Fleming, P. Jamison, J. Brown, and R. Arndt, "80nm Poly-Silicon Gated n-FETs with Ultra-Thin Al_2O_3 Gate Dielectric for ULSI Applications," *IEDM Tech. Digest*, pp. 223–226 (2000).
36. J. W. Klaus and S. M. George, "Atomic Layer Deposition of Tungsten Using Sequential Surface Chemistry with Sacrificial Stripping Reaction," *Thin Solid Films* **360,** 145–153 (2000).
37. D.-J. Kim, Y.-B. Jung, M.-B. Lee, Y.-H. Lee, J.-H. Lee, and J.-H. Lee, "Applicability of ALE TiN Films as Cu/Si Diffusion Barriers," *Thin Solid Films* **372,** 276–283 (2000).
38. J. W. Klaus and S. M. George, "Atomic Layer Deposition of Tungsten Nitride Films Using Sequential Surface

Reactions," *J. Electrochem. Soc.* **147,** 1175–1181 (2000).
39. S. Hirose, A. Yoshida, M. Yamaura, K. Hara, and H. Munekata, "Selective Area Growth by Metal Organic Vapor Phase Epitaxy and Atomic Layer Epitaxy Using Ga_2O_3 as a Novel Mask Layer," *Jpn. J. Appl. Phys. Part 1* **38,** 1516–1520 (1999).
40. H. Isshiki, Y. Aoyagi, T. Sugano, S. Iwai, and T. Meguro, "Crystallographic Selective Growth of GaAs by Atomic Layer Epitaxy," *Appl. Phys. Lett.* **63,** 1528–1530 (1993).
41. D. J. Economou and R. C. Alkire, "Effect of Potential Field on Ion Deflection and Shape Evolution of Trenches During Plasma Assisted Etching," *J. Electrochem. Soc.* **135,** 941–949 (1988).
42. J. W. Coburn and H. F. Winters, "Conductance Considerations in the Reactive Ion Etching of High Aspect Ratio Features," *Appl. Phys. Lett.* **55,** 2730–2732 (1989).
43. B. E. Thompson, H. H. Sawin, and D. A. Fischer, "Monte Carlo Simulation of Ion Transport Through RF Glow-Discharge Sheaths," *J. Appl. Phys.* **63,** 2241–2251 (1988).
44. K. Nojiri, E. Iguchi, K. Kawamura, and K. Kadota, in *Extended Abstracts*, *21st Conference on Solid State Devices and Materials*, Business Center for the Academic Society of Japan, Tokyo, 1989, pp. 153–156.
45. K. Suzuki, S. Hiraoka, and S. Nishimatsu, "Anisotropic Etching of Polycrystalline Silicon with Hot Cl_2 Molecular Beam," *J. Appl. Phys.* **64,** 3697–3705 (1988).
46. T. Ono, H. Kashima, S. Hiraoka, and K. Suzuki, "Anisotropic Etching of GaAs Using a Hot Cl_2 Molecular Beam," *J. Vac. Sci. Technol. B* **9,** 2798–2801 (1991).
47. M. W. Geis, N. N. Efremow, S. W. Pang, and A. C. Anderson, "Hot Jet Etching of Pb, GaAs, and Si," *J. Vac. Sci. Technol. B* **5,** 363–365 (1987).
48. Y. Teraoka and I. Nishiyama, "Anisotropic Si(100) Etching Induced by High Translational Energy Cl_2 Molecular Beams," *Appl. Phys. Lett.* **63,** 3355–3357 (1993).
49. K. P. Giapis, T. A. Moore, and T. K. Minton, "Hyperthermal Neutral Beam Etching," *J. Vac. Sci. Technol. A* **13,** 959–965 (1995).
50. G. S. Hwang, C. M. Anderson, M. J. Gordon, T. A. Moore, T. A. Minton, and K. P. Giapis, "Gas-Surface Dynamics and Profile Evolution During Etching of Silicon," *Phys. Rev. Lett.* **77,** 3049–3052 (1996).
51. F. Shimokawa, H. Tanaka, Y. Uenishi, and R. Sawada, "Reactive-Fast-Atom Beam Etching of GaAs Using Cl_2 Gas," *J. Appl. Phys.* **66,** 2613–2618 (1989).
52. T. Mizutani and S. Nishimatsu, "Sputtering Yield and Radiation Damage by Neutral Beam Bombardment," *J. Vac. Sci. Technol. A* **6,** 1417–1420 (1988).
53. T. Tsuchizawa, Y. Jin, and S. Matsuo, "Generation of Electron Cyclotron Resonance Neutral Stream and Its Application to Si Etching," *Jpn. J. Appl. Phys. Part 1* **33,** 2200–2206 (1994).
54. M. J. Groeckner, T. K. Bennett, and S. A. Cohen, "A Source of Hyperthermal Neutrals for Materials Processing," *Appl. Phys. Lett.* **71,** 980–982 (1997).
55. K. Yokogawa, T. Yunogami, and T. Mizutani, "Neutral-Beam-Assisted Etching System for Low-Damage SiO_2 Etching of 8-Inch Wafers," *Jpn. J. Appl. Phys. Part 1* **35,** 1901–1905 (1996).
56. T. Meguro, M. Hamagaki, S. Modaressi, T. Hara, Y. Aoyagi, M. Ishii, and Y. Yamamoto, "Digital Etching of GaAs: New Approach of Dry Etching to Atomic Ordered Processing," *Appl. Phys. Lett.* **56,** 1552–1554 (1990).
57. S. D. Athavale and D. J. Economou, "Molecular Dynamics Simulation of Atomic Layer Etching of Silicon," *J. Vac. Sci. Technol. A* **13,** 966–971 (1995).
58. M. Ishii, T. Meguro, H. Kodama, Y. Yamamoto, and Y. Aoyagi, "Study of Surface Processes in the Digital Etching of GaAs," *Jpn. J. Appl. Phys.* **31,** 2212–2215 (1992).
59. W. M. Lau, L. J. Huang, W. H. Chang, M. Vos, and I. V. Mitchell, "Ultrashallow Depth Profiling Using Ozone Oxidation and HF Etching of Silicon," *Appl. Phys. Lett.* **63,** 78–80 (1993).
60. S. J. Jeng, W. C. Natzel, and C. Yu, "Device and Method for Accurate Etching and Removal of Thin Film," U.S. Patent 5,282,925, filed 1996, issued 1998.
61. C. Hu, D. C. Ahlgren, E. G. Walton, G. B. Bronner, and W. C. Natzel, "Oxide Strip That Improves Planarity," U.S. Patent 5,766,971, filed 1996, issued 1998.
62. L. Chang, S. Tang, T.-J. King, J. Bokor, and C. Hu, "Gate Length Scaling and Threshold Voltage Control of Double-Gate MOSFETs," *IEDM Tech. Digest*, pp. 719–722 (2000).
63. P. D. Agnello, T. O. Sedgwick, and J. Cotte, "Growth Rate Enhancement of Heavy n- and p-Type Doped Silicon Deposited by Atmospheric-Pressure Chemical Vapor Deposition at Low Temperatures," *J. Electrochem. Soc.* **140,** 2703–2709 (1993).
64. Y. Mitani, I. Mizushima, S. Kambyashi, H. Koyama, M. T. Takagi, and M. Kashiwagi, "Buried Source and Drain (BSD) Structure for Ultra-Shallow Junction Using Selective Deposition of Highly Doped Amorphous Silicon," *Symposium on VLSI Technology*, *Digest of Technical Papers*, 1996, pp. 176–177.
65. M. Ieong, P. M. Solomon, S. E. Laux, H.-S. P. Wong, and D. Chidambarro, "Comparison of Raised and Schottky Source/Drain MOSFETs Using a Novel Tunneling Contact Model," *IEDM Tech. Digest*, pp. 733–736 (2000).
66. J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T. J. King, and C. Hu, "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20nm Gate Length Regime," *IEDM Tech. Digest*, pp. 57–60 (2000).
67. J. W. Matthews and A. E. Blakeslee, "Defects in Epitaxial Multilayers," *J. Cryst. Growth* **27,** 118–126 (1974).
68. R. People and J. Bean, "Calculation of Critical Layer Thickness Versus Lattice Mismatch for $Ge_{1-x}Si_x$/Si Strained-Layer Heterostructure," *Appl. Phys. Lett.* **47,** 322 (1985).
69. H. Kressel, "The Application of Heterojunction Structures to Optical Devices," *J. Electron. Mater.* **4,** 1081 (1975).
70. Y. H. Lo, "New Approach to Grow Pseudomorphic Structures Over the Critical Thickness," *Appl. Phys. Lett.* **59,** 2311–2313 (1991).
71. D. Teng and Y. H. Lo, "Dynamic Model for Pseudomorphic Structures Grown on Compliant Substrates: An Approach to Extend the Critical Thickness," *Appl. Phys. Lett.* **62,** 43–45 (1993).
72. C. L. Chua, W. Y. Hsu, C. H. Lin, G. Christenson, and Y. H. Lo, "Overcoming the Pseudomorphic Critical Thickness Limit Using Compliant Substrates," *Appl. Phys. Lett.* **64,** 3640–3642 (1994).
73. A. R. Powell, S. S. Iyer, and F. K. LeGoues, "New Approach to the Growth of Low Dislocation Relaxed SiGe Material," *Appl. Phys. Lett.* **64,** 1856–1858 (1994).
74. F. K. LeGoues, A. R. Powell, and S. S. Iyer, "Relaxation of SiGe Thin Films Grown on Si/SiO_2 Substrates," *J. Appl. Phys.* **75,** 7240–7246 (1994).
75. Z. Yang, J. Alperin, W. I. Wang, S. S. Iyer, T. S. Kuan, and F. Semendy, "In Situ Relaxed $Si_{1-x}Ge_x$ Epitaxial Layers with Low Threading Dislocation Densities Grown on Compliant Si-on-Insulator Substrates," *J. Vac. Sci. Technol. B* **16,** 1489–1491 (1998).
76. Y. H. Luo, J. L. Liu, G. Jin, J. Wan, K. L. Wang, C. D. Moore, M. S. Goorsky, C. Chih, and K. N. Tu, "Effective Compliant Substrate for Low-Dislocation Relaxed SiGe Growth," *Appl. Phys. Lett.* **78,** 1219–1221 (2001).
77. F. Y. Huang, M. A. Chu, M. O. Tanner, K. L. Wang, G. D. U'Ren, and M. S. Goorsky, "High-Quality Strain-Relaxed SiGe Alloy Grown on Implanted Silicon-on-Insulator Substrate," *Appl. Phys. Lett.* **76,** 2680–2682 (2000).

78. C. Carter-Coman, A. S. Brown, N. M. Jokerst, D. E. Dawson, R. Bicknell-Tassius, Z. C. Feng, K. C. Rajkumar, and G. Dagnall, "Strain Accommodation in Mismatched Layers by Molecular Beam Epitaxy: Introduction of a New Compliant Substrate Technology," *J. Electron. Mater.* **25,** 1044–1048 (1996).
79. S. I. Romanov, V. I. Madhanov, L. V. Sokolov, A. Gutakovskii, and O. P. Pchelyakov, "GeSi Films with Reduced Dislocation Density Grown by Molecular-Beam Epitaxy on Compliant Substrates Based on Porous Silicon," *Appl. Phys. Lett.* **76,** 4118–4120 (1999).
80. F. E. Ejeckam, Y. H. Lo, S. Subramanian, H. Q. Hou, and B. E. Hammons, "Lattice Engineered Compliant Substrate for Defect Free Heteroepitaxial Growth," *Appl. Phys. Lett.* **70,** 1685–1687 (1997).
81. F. E. Ejeckam, M. L. Seaford, Y. H. Lo, H. Q. Hou, and B. E. Hammons, "Dislocation-Free InSb Grown on GaAs Compliant Universal Substrates," *Appl. Phys. Lett.* **71,** 776–778 (1997).
82. J. Cao, D. Pavlidis, A. Eisenbach, A. Philipe, C. Bru-Chevallier, and G. Guillot, "Photoluminescence Properties of GaN Grown on Compliant Silicon-on-Insulator Substrates," *Appl. Phys. Lett.* **71,** 3880–3882 (1997).
83. Z. Yang, F. Guarin, I. W. Tao, W. I. Wang, and S. Iyer, "Approach to Obtain High Quality GaN on Si and SiC-on-Insulator Compliant Substrate by Molecular-Beam Epitaxy," *J. Vac. Sci. Technol. B* **13,** 789–791 (1995).
84. A. S. Brown, "Compliant Substrate Technology; Status and Prospects," *J. Vac. Sci. Technol. B* **16,** 2308–2312 (1998).
85. G. Kastner and U. Gosele, "Principles of Strain Relaxation in Heteroepitaxial Films Growing on Compliant Substrates," *J. Appl. Phys.* **88,** 4048–4055 (2000).
86. P. D. Moran, D. M. Hansen, R. J. Matyi, L. J. Mawst, and T. F. Kuech, "Experimental Test for Elastic Compliance During Growth of Glass-Bonded Compliant Substrates," *Appl. Phys. Lett.* **76,** 2541–2543 (2000).

Received June 16, 2001; accepted for publication January 19, 2002

Paul D. Agnello *IBM Research Division, Semiconductor Research and Development Center, Hopewell Junction, New York 12533 (agnellop@us.ibm.com).* Dr. Agnello received his B.S., M.S., and Ph.D. degrees, all from the Electrical, Computer and Systems Engineering Department of Rensselaer Polytechnic Institute. He joined the IBM Thomas J. Watson Research Center in 1988. Dr. Agnello has received IBM Outstanding Technical Achievement Awards for low-temperature CVD processing and alloy silicide development. In addition, he has received an IBM Outstanding Innovation Award for the development of 0.25-μm CMOS technology and a Corporate Award for high-performance CMOS logic development. Recently he led the development of 0.18-μm-generation Cu back-end-of-line for high-performance CMOS; he currently manages the high-performance CMOS integration project responsible for 0.13-μm and beyond silicon-on-insulator technologies. Dr. Agnello is the author or co-author of more than 60 publications; he holds 15 U.S. patents.